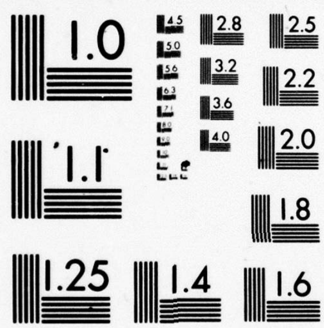
AD-A062 177 MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTE-ETC F/G 20/6
REFLECTANCE MAP TECHNIQUES FOR ANALYZING SURFACE DEFECTS--ETC.(U)
JUN 78 R. J. WOODHAM N00014-77-C-0389
UNCLASSIFIED AI-TR-457 N/L

1 OF 3
AD A
062177



MICROCOPY RESOLUTION TEST CHART

AI-TR-457

Technical rept.

Reflectance Map Techniques for Analyzing Surface Defects in Metal Castings

N00014-77-C-0389, N00014-75-C-0643

220p.

Robert J. Woodham

Jun[redacted]78

ARTIFICIAL INTELLIGENCE LABORATORY

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

407 483

220

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER TR-457	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Reflectance Map Techniques for Analyzing Surface Defects in Metal Castings		5. TYPE OF REPORT & PERIOD COVERED technical report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Robert J. Woodham		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0643 N00014-77-C-0389
9. PERFORMING ORGANIZATION NAME AND ADDRESS Artificial Intelligence Laboratory 545 Technology Square Cambridge, Massachusetts 02139		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Advanced Research Projects Agency 1400 Wilson Blvd Arlington, Virginia 22209		12. REPORT DATE June 1978
		13. NUMBER OF PAGES 216
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Office of Naval Research Information Systems Arlington, Virginia 22217		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Image Analysis
Image Understanding
Reflectance Map
Photogrammetry
Photometric Stereo
Automatic Inspection
Surface Curvature
Image Hessian

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report explores the relation between image intensity and object shape. It is shown that image intensity is related to surface orientation and that a variation in image intensity is related to surface curvature. Computational methods are developed which use the measured intensity variation across surfaces of smooth objects to determine surface orientation.

In general, surface orientation is not determined locally by the intensity value recorded at each image point. Tools are needed to explore the problem

of determining surface orientation from image intensity. The notion of gradient space, popularized by Huffman and Mackworth, is used to represent surface orientation. The notion of a reflectance map, originated by Horn, is used to represent the relation between surface orientation and image intensity. The image Hessian is defined and used to represent surface curvature. Properties of surface curvature are expressed as constraints on possible surface orientations corresponding to a given image point. Methods are presented which embed assumptions about surface curvature in algorithms for determining surface orientation from the intensities recorded in a single view.

If additional images of the same object are obtained by varying the direction of incident illumination, then surface orientation is determined locally by the intensity values recorded at each image point. This fact is exploited in a new technique called photometric stereo.

The visual inspection of surface defects in metal castings is considered. Two casting applications are discussed. The first is the precision investment casting of turbine blades and vanes for aircraft jet engines. In this application, grain size is an important process variable. The existing industry standard for estimating the average grain size of metals is implemented and demonstrated on a sample turbine vane. Grain size can be computed from the measurements obtained in an image, once the foreshortening effects of surface curvature are accounted for. The second is the green sand mold casting of shuttle eyes for textile looms. Here, physical constraints inherent to the casting porcess translate into constraints on the surface topography of cast objects. In order to exploit these constraints, it is necessary to interpret features of intensity as features of object shape.

Both applications demonstrate that successful visual inspection requires the ability to interpret observed changes in intensity in the context of surface topography. The theoretical tools developed in this report provide a framework for this interpretation.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research was provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research Contract N00014-75-C-0643 and in part by the Office of Naval Research under Office of Naval Research Contract N00014-77-C-0389.

REFLECTANCE MAP TECHNIQUES FOR ANALYZING SURFACE DEFECTS IN METAL CASTINGS

by

ROBERT J. WOODHAM

Massachusetts Institute of Technology

June 1978

D.H

Revised version of a dissertation submitted to the Department of Electrical Engineering and Computer Science on September 7, 1977 in partial fulfillment of the requirements of the Degree of Doctor of Philosophy.

ABSTRACT

This report explores the relation between image intensity and object shape. It is shown that image intensity is related to surface orientation and that a variation in image intensity is related to surface curvature. Computational methods are developed which use the measured intensity variation across surfaces of smooth objects to determine surface orientation.

In general, surface orientation is not determined locally by the intensity value recorded at each image point. Tools are needed to explore the problem of determining surface orientation from image intensity. The notion of gradient space, popularized by Huffman and Mackworth, is used to represent surface orientation. The notion of a reflectance map, originated by Horn, is used to represent the relation between surface orientation and image intensity. The image Hessian is defined and used to represent surface curvature. Properties of surface curvature are expressed as constraints on possible surface orientations corresponding to a given image point. Methods are presented which embed assumptions about surface curvature in algorithms for determining surface orientation from the intensities recorded in a single view.

If additional images of the same object are obtained by varying the direction of incident illumination, then surface orientation is determined locally by the intensity values recorded at each image point. This fact is exploited in a new technique called photometric stereo.

The visual inspection of surface defects in metal castings is considered. Two casting applications are discussed. The first is the precision investment casting of turbine blades and vanes for aircraft jet engines. In this application, grain size is an important process variable. The existing industry standard for estimating the average grain size of metals is implemented and demonstrated on a sample turbine vane. Grain size can be computed from the measurements obtained in an image, once the foreshortening effects of surface curvature are accounted for. The second is the green sand mold casting of shuttle eyes for textile looms. Here, physical constraints inherent to the casting process translate into constraints on the surface topography of cast objects. In order to exploit these constraints, it is necessary to interpret features of intensity as features of object shape.

Both applications demonstrate that successful visual inspection requires the ability to interpret observed changes in intensity in the context of surface topography. The theoretical tools developed in this report provide a framework for this interpretation.

Thesis Supervisor: Berthold K. P. Horn

Title: Associate Professor of Electrical Engineering and Computer Science

ACKNOWLEDGEMENTS

I would like to thank my thesis supervisor, Professor Berthold Horn, and my thesis readers, Professor Patrick Winston and Professor David Marr, for their support and enthusiasm throughout.

I would also like to thank my two friends and colleagues Mark Lavin and Tomas Lozano-Perez of the M.I.T. AI Lab who were always willing to contribute ideas both of a technical nature and other.

Dr. Mike Fassler, Dr. John Erickson and Dr. Pat Sullivan of the Casting Development Section, Materials Engineering & Research Laboratory, Pratt & Whitney Aircraft and Dr. Jim Blackler of the Draper Division Foundry, Rockwell International provided useful advice and technical assistance.

Anna R. Bruss helped proofread and improve technical style through several drafts of this report.

This research was made possible because of the unique facilities and atmosphere of the M.I.T. Artificial Intelligence Laboratory. For this I have many students, professors, hackers and friends, both past and present, to thank.

I am grateful for early financial support received from the Woodrow Wilson Foundation and the Canada Council.

Finally, a personal note of gratitude to two special people - Louise England, lawyer, and Dr. Suzanne Lacasse, civil engineer, who have always been there to bail me out of trouble and keep my foundations secure.

TABLE OF CONTENTS

1. INTRODUCTION

This report explores the relation between image intensity and object shape. The research strategy is to exploit, as directly as possible, the intensity values recorded in an image. In order to extract the information contained in image intensity, it is necessary to understand both the geometry of image projection and the radiometry of image formation. It is shown that image intensity is related to surface orientation and that a variation in image intensity is related to surface curvature.

Basic tools are needed. Gradient space, popularized in [Huffman 71] and [Mackworth 73], is used to represent surface orientation. The reflectance map, introduced in [Horn 77], is used to represent the relation between surface orientation and image intensity. With these tools, a fixed scene illumination, object photometry and imaging geometry can be incorporated into an explicit model that allows image intensity to be related directly to surface orientation. This relationship is not functional since surface orientation has two degrees of freedom and image intensity provides only one measurement.

Nevertheless, properties of surface curvature can be expressed as constraints on possible surface orientations corresponding to a given image point. The image Hessian is defined and used to represent surface curvature. Computational methods are presented which embed assumptions about surface curvature in algorithms for determining surface orientation from the intensities recorded in a single view.

If additional images of the same object are obtained by varying the direction of incident illumination, then surface orientation is determined locally by the intensity values recorded at each image point. This fact is exploited in a new technique called photometric stereo.

Interpreting image intensity in terms of underlying object shape is the key to the automatic visual inspection of metal castings. Two casting applications are discussed. The first is the precision investment casting of turbine blades and vanes for aircraft jet engines. In this application, grain size is an important process variable. The existing industry standard for estimating the average grain size of metals is implemented and demonstrated on a sample turbine vane. Grain size can be computed from the measurements obtained in an image provided that the foreshortening effects of surface curvature can *be* accounted for. *Foreshortening* can be accounted for once the underlying object shape is known.

The second is the green sand mold casting of shuttle eyes for textile looms. Here, physical constraints inherent to the casting process translate into constraints on the surface topography of cast objects. In order to exploit these constraints, however, it is necessary to interpret features of intensity as features of object shape.

Both applications demonstrate that successful visual inspection requires the ability to interpret observed changes in intensity in the *context of surface topography*. The theoretical tools developed in this report provide a framework for this interpretation.

1.1 IMAGE ANALYSIS VERSUS SCENE ANALYSIS

Vision is a tough problem. One thing that seems to make vision such a tough problem is the fact that many different kinds of knowledge can influence the interpretation of an image. An important question to ask concerns how much of the interpretation of an image is forced by the data itself and how much can be attributed to the influence of prior *expectation*. This is a difficult question to deal with in human perception

because it is impossible to factor out experimentally one effect from the other. At best, one can characterize human vision as the interaction of diverse, partially complete and possibly redundant knowledge sources.

In machine vision, the distinction between "data forced" and "prior expectation" can be dealt with by considering vision to be a two stage process. One begins with the intensity values recorded in an image. *Image analysis* extracts features from the raw intensity values and converts these features into a convenient symbolic representation. *Scene analysis* interprets the symbolic features produced by image analysis according to some externally defined goal. Image analysis defines what can be considered as forced by the data in the subsequent interpretation. Scene analysis, on the other hand, is an exercise in problem solving. In scene analysis, one is free to invoke whatever prior knowledge is available to aid in image interpretation.

Early artificial intelligence research in machine vision concentrated on images of scenes containing plane-faced polyhedra. Initially, the distinction between image analysis and scene analysis seemed clear. The purpose of image analysis was to generate a two-dimensional line drawing of the scene [Binford and Horn 73]. The purpose of scene analysis was to interpret a two-dimensional line drawing in terms of the three-dimensional objects which gave rise to it [Roberts 65], [Guzman 68], [Huffman 71], [Clowes 71], [Mackworth 73], [Waltz 75], [Winston 75].

As the field matured, the actual distinction between image analysis and scene analysis became less clear. A richer form of interaction between image analysis and scene analysis was achieved [Falk 72], [Winston 73], [Freuder 76]. One benefit of this new interaction was a reduction in the computation required in image analysis. Sensitive line-finding procedures

existed but generated too many false targets to be applied uniformly over an image. Nevertheless, these procedures could be effectively applied to verify the presence of lines hypothesized by scene analysis routines [Shirai 75].

A conceptual distinction between image analysis and scene analysis persists. A prevailing notion is that the nature of the features extracted from an image is of less consequence than the subsequent interpretation of those features. The failings of image analysis are to be compensated for by "smarter" scene analysis. Research is lead away from basic image analysis towards more elaborate scene analysis. Avoiding image analysis is often not so much a question of research philosophy as it is one of practical necessity. Yet, it seems critical to determine what information can be extracted in image analysis before attempting to make a distinction between image analysis and scene analysis for a particular problem.

Recent work by [Marr 76] and [Horn 77] has demonstrated that there is a great deal of information about three-dimensional shape contained in image intensities and that this information can be computed without recourse to higher-level knowledge. [Horn 77], for example, shows how a careful analysis of the intensity profiles in images of polyhedra can be used to interpret certain two-dimensional lines directly as convex, concave or occluding edges. Adding this information directly to a line drawing produced in image analysis would make much subsequent scene analysis superfluous.

1.2 IMAGE ANALYSIS IS HARD

Much of the work in machine vision does not explicitly exploit the information contained in the intensity values recorded in an image. In this section, reasons are suggested why image analysis is hard. The purpose is to point out some of the difficulties associated with interpreting image intensities. Only by taking cognizance of these difficulties is it possible to understand when and why various machine vision techniques will work and when and why they will fail.

1.2.1 THERE IS A LOT OF DATA IN AN IMAGE

One of the major stumbling blocks to image analysis is the sheer quantity of data present in an image. One goal of image analysis is to extract features of intensity that are important and to throw everything else away. Image analysis often becomes an exercise in data compression. But, the features one can use in a data compression process are those which can be conveniently defined in terms of properties of images. There is not always a simple correlation between properties of images and properties of the objects which give rise to those images.

As a rough rule of thumb, one can say that practical vision systems exist only for domains which have the following two properties:

- There is a simple correlation between properties of interest in the domain and properties of images of the domain.
- These image properties are simple to compute.

A good correlation between properties of the domain and properties of images of the domain occurs in domains that are inherently two-dimensional. A significant reduction both in the quantity of data present in an image

and in the complexity of feature computation occurs in domains in which intensity can be considered to have only two values. Typical applications of machine vision systems involve images, often binary, of two-dimensional worlds. Examples are: optical character recognition (OCR); blood cell analysis [*Young* 69]; detection of etching defects in printed circuit boards [Ejiri *et al* 73].

1.2.2 IMAGE PROJECTION CAUSES PROBLEMS

Images are defined in two dimensions while objects exist in a three-dimensional world. Information is lost in the projection of a three-dimensional object onto a two-dimensional image. The mapping from object space to image space is many-to-one. It is impossible to analyze two-dimensional images without specific assumptions about the three-dimensional nature of the objects that gave rise to them. A particular object feature may appear quite differently depending on the viewing direction. More insidious is the fact that projection introduces two-dimensional image features which have no direct correlation with any three-dimensional object property. Neighboring points in an image do not necessarily correspond to neighboring points on objects. Parts of one object may obscure parts of another.

1.2.3 IMAGE ILLUMINATION CAUSES PROBLEMS

The same object shape viewed with the same imaging geometry will generate different image intensities depending on the direction and nature of the incident illumination. Changes in illumination may cause a particular object feature to appear quite differently even when seen from the same viewing direction.

1.2.4 SURFACE PHOTOMETRY CAUSES PROBLEMS

The same object shape viewed with the same imaging and light source geometry will generate different image intensities depending on the surface photometry of the object material. Surface photometry varies from object material to object material. For a fixed object material, surface photometry varies depending on whether the object is wet or dry, clean or dirty.

1.2.5 THE HUMAN VISUAL SYSTEM IS REMARKABLY FORGIVING

The human visual system does a remarkable job of interpreting image intensity despite problems caused by projection, illumination and surface photometry outlined above. At first glance, this might be cited as evidence that image analysis cannot intrinsically be hard. This is false optimism and soon leads to serious difficulty.

Designing an image analysis system by letting intuition decide what the intensities ought to be guarantees failure. It is necessary to obtain digital images and to examine the actual numbers recorded. More importantly, image analysis requires good physical models of how surfaces reflect light.

1.3 INSPECTION IS A GOOD DOMAIN FOR IMAGE ANALYSIS

Normal visual acuity does not guarantee success as an inspector. The initial problem in inspection is to learn how to interpret the visual data. Then, however, the problem becomes one of developing an ablility to eliminate the influence of prior expectation on visual interpretation. Inspection can be likened to proofreading text. If one reads sentences, the expectation of what a word should be makes it very difficult to catch

typographical errors. To proofread effectively, it is necessary to ignore the semantic context of what one is reading and concentrate instead on the individual words and letters. Thus, an inspector is trained to see what he is *forced* to see rather than what he expects to see. This is precisely what makes inspection a difficult job for people. This is also what makes inspection the right job to hand over to machine vision systems.

In machine vision, the influence of prior expectation can be carefully *controlled*. In visual inspection, it is necessary to rely on the information contained in the actual intensity values present in an image. This makes inspection a good domain for exploring what can and cannot be determined from intensity data alone. Inspection forces one to deal explicitly with the issue of what interpretation is forced by the data.

1.4 METAL CASTING IS A GOOD DOMAIN FOR INSPECTION

There are important reasons why metal casting is a good domain for research in automatic inspection. One set of reasons relates to the demand for automatic inspection systems:

- A foundry is a hostile work environment. OSHA and EPA regulations are forcing foundries to modernize their equipment and safety practices. But making foundries more acceptable to humans is costly. A better idea is to automate humans out of hazardous areas. A formidable stumbling block to the automatic foundry is the need for inspection.
- Strong foreign competition is forcing the casting industry to re-evaluate its materials usage and energy costs. The goal is to minimize overdesign and associated

materials waste by achieving tighter control over the manufacturing process. Tighter control requires more inspection.

- Productivity in casting inspection is low. Maintaining tight quality standards is a difficult and costly operation. Inspection is a tedious job for humans. Tasks requiring close visual attention create fatigue. Repetitive tasks induce boredom. Yet inspection demands mental and physical alertness. Automation is the key to increasing inspection productivity.
- Casting is a vital component of the metal working industry. The trend is towards the increased importance of casting as a primary fabrication process. More sophisticated casting reduces the number of expensive machining operations required to complete the part. More dependable casting reduces the number of expensive machining operations wasted on defective parts. Sophisticated and dependable casting requires better quality inspection.

Another set of reasons for why metal casting is a good domain for research in automatic inspection relates to the need for flexible, computer-based systems:

- Foundries are job-shop environments. The typical foundry casts many different part geometries in small to medium production quantities. The specialized techniques available for the automation of large volume production are not suited to the foundry. The palletized, pick and

place, orientation preserving style of parts handling envisioned by most automation engineers is the antithesis of the parts handling problem in the foundry. Castings are tumbled freely in common shaker machines to break away residual mold material and gating. Castings are cleaned in common sand-blast machines. It is not feasible to redesign these operations to preserve part identity and orientation. On-line inspection must be flexible enough to accommodate different part geometries and different methods of part presentation.

- The inspection of a casting does not result in a simple PASS/FAIL decision. Inspection requires interpretation. Often, this interpretation depends on different kinds of knowledge. Knowledge of the casting process determines the nature of the defects to be found. Knowledge of subsequent machining operations determines the acceptability of surface imperfections. Knowledge of in-service stress patterns determines the critical tolerances for each subsection of the part. In order to apply existing standards, automatic inspection systems must be able to interpret test data within the context of the part as a whole.
- The inspection of castings remains something of an art. Often, there is only a loose correlation between test results and hard evidence as to what functional characteristics of the part are actually being measured. Current research in nondestructive testing emphasizes the

extraction of more information from existing test data as much as it does the development of new test techniques. The ability to store and retrieve test data on a computer system adds a new dimension to nondestructive testing. Computer-based interpretation is the key to exploiting as much of the information in the test data as possible.

- Foundries depend critically on the experience of their inspectors. One consequence of this dependence is that there is little or no industry-wide standardization of test interpretation. There is a strong desire to develop such standards for casting inspection. Automation of test interpretation is one way to achieve standardization. Flexibility in automation is the key to making such standardization useful.
- There is no casting free of imperfections. Imperfections are inherent to the casting process. The goal of manufacturing is to hold these imperfections to specified tolerances. In the automation of casting inspection, the emphasis must be away from simple detection devices and towards computer-based systems. The goal is to allow maximum flexibility in the specification and interpretation of inspection standards.

1.5 THE RESEARCH APPROACH

Several questions require attention:

- How are images formed?
- How does the real world constrain an image?
- How is it possible to exploit these constraints in order to interpret the intensity values recorded in an image?

One can begin by writing down an equation describing the image-forming process as follows:

<image intensity> = <incident illumination> x <surface reflectance> (1.5.1)

The terms image intensity, incident illumination and surface reflectance must be made precise, but for the moment, the technical details will be omitted in order to make some qualitative observations. *In writing* down this equation, emphasis is being placed on the fact that image intensity can be treated as a measurement arising from a real physical process. Images can be interpreted by understanding the underlying physical process which gave rise to them.

Four factors interact in image formation:

- imaging geometry
- incident illumination
- surface photometry
- surface topography

The imaging geometry determines the projection of the three-dimensional object space coordinates onto two-dimensional image coordinates. Without illumination, there can be no image. Incident illumination is characterized by the spatial and spectral distribution and state of polarization of radiant energy falling on the object surface.

Surface photometry determines how light reflects off the object surface. It is determined by the optical constants of the object material and by the surface microstructure. (Surface microstructure is surface detail which is too fine to be resolved in the image but which causes observable effects in the way light is reflected at the object surface.) Surface topography is surface detail which is within the resolution limits of the imaging hardware. It characterizes the gross object shape relative to the viewer.

Equation (1.5.1) relates these four factors. The notions of gradient space and the reflectance map allow this equation to be made precise. For fixed surface photometry, incident illumination and imaging geometry, this equation is used to study how to determine surface topography from image intensity.

The work presented in this report is based on a physical model of how images are formed. No attempt is made to account for the subjective phenomena associated with human perception. [Beck 74], for example, contains an image of a "matte" vase. By adding two local "specularities" to this image, a new picture is created in which the entire vase appears shiny. Regions in the first image appear dull while the same regions in the second image appear shiny, even though the intensity values recorded are identical. Equation (1.5.1) accounts only for the intensity values recorded in an image and not for the human perception of those values.

1.5.1 RELATED APPLICATIONS

There are a variety of uses for equation (1.5.1). In computer graphics, one problem is the generation of gray-level images from surface models. Using an assumed surface photometry, incident illumination and imaging geometry, (1.5.1) can be used to synthesize images of objects.

Since gray-level images are readily interpreted by humans, this is a useful tool in such diverse fields as computer aided design and automated cartography. Work in computer graphics has led to a number of phenomenological models of surface reflectance [Phong 75].

If the surface topography, incident illumination and imaging geometry are known, then (1.5.1) can be used to determine surface photometry from image intensity. If the spectral nature of incident illumination or the spectral response of the imaging hardware is varied, then the dependence of surface photometry on the wavelength of incident illumination can be measured. This has been applied to determining ground cover from multispectral LANDSAT imagery registered to digital terrain models [Horn & Bachman 77].

If the surface topography, incident illumination, imaging geometry and surface photometry are known, then (1.5.1) can be used either to determine the optical constants of the object material, assuming a known surface microstructure, or to determine surface microstructure, assuming known optical constants. Reflectance spectroscopy uses (1.5.1) to determine optical constants. This field has become an important new tool in analytic chemistry. There are many materials that cannot be analyzed by traditional spectroscopic methods. Some of these materials, however, can be ground into fine powders of known particle size and shape. In turn, analytic models have been developed to relate the surface photometry of such powders to the optical constants of the object material of which they are composed [Wendlandt & Hecht 66].

Another example analyzes how the observed pattern of sunlight (or moonlight) glitter on a wind-ruffled sea can be used to deduce information about the sea state [Plass *et al* 77]. This technique is proposed both as a

method for evaluating models of ocean wave structure and for making measurements of the parameters that determine sea roughness, such as wind speed and direction. Analyzing the reflection of light from a flat water surface is a simple exercise in the geometry of specular reflection. Waves introduce surface microstructure. The observed glitter pattern consists of numerous instantaneous "glints" produced by perfect specular reflection from wave facets having the appropriate momentary orientation. The shape of the glitter pattern is determined by the distribution of wave inclinations on the water surface. Determining this distribution requires that the light source and imaging geometry be known since these introduce modifications in the apparent wave-slope distribution.

The above examples illustrate the importance of explicitly accounting for the dependence of image intensity on the four factors: surface photometry, surface topography, incident illumination and imaging geometry. In most applications, it is convenient to standardize one or more of these factors in order to determine the dependence of image intensity on the others.

1.6 SUMMARY OF THE REPORT

Chapter 1 introduces the research. A distinction between image analysis and scene analysis is made. Reasons are suggested for why image analysis is hard. Nevertheless, image analysis can benefit from an explicit model of the image-forming process. Such a model must account for the effects of surface photometry, surface topography, incident illumination and imaging geometry. Automatic inspection is presented as a good domain for image analysis. Metal casting is presented as a good domain for automatic inspection.

Chapter 2 introduces a framework for interpreting the intensity values recorded in an image. Image analysis is formulated as the problem of determining, at each image point, the range, the surface orientation and the surface curvature at the corresponding object point. Viewer-centered representations for surface orientation and surface curvature are developed. Gradient space is used to represent surface orientation. The image Hessian matrix is used to represent surface curvature. The reflectance map establishes the relationship between surface orientation and image intensity. The image Hessian matrix establishes the relationship between movement in the image and change in surface orientation. Image intensity constrains the image Hessian matrix but does not determine it completely. Assumptions about surface curvature further constrain the Hessian matrix. These results are demonstrated in a particular relaxation algorithm for determining surface orientation from a single view. Propagation of local constraint over regions of the image is used to assign a global interpretation to the surface. The algorithm is illustrated using a synthesized image of a sphere and the assumption that it is a convex solid of revolution.

Chapter 3 develops the relationship between surface curvature and image intensity in more detail. Known properties of surface curvature can simplify image analysis. A number of cases are considered. These are surfaces with constant image Hessian, surfaces that are singly curved and surfaces of generalized cones. These special cases illustrate when shape information can be determined solely from image intensity and when shape information can be determined only when additional information is provided, either implicitly, in the form of assumptions about surface curvature, or explicity from other sources, such as object boundaries. Finally, if

additional images of the same object are obtained by varying the direction of incident illumination, then surface orientation can be determined locally by the intensity values recorded at each image point. This fact is exploited in a new technique called photometric stereo.

Chapter 4 provides a brief introduction to the problem of metal casting inspection. Two particular casting applications are presented. One is the precision investment casting of turbine blades and vanes for aircraft jet engines. The other is the green sand mold casting of shuttle eyes for textile looms. This chapter is not critical to the technical development. Rather, it serves as a general introduction to casting inspection for the non metallurgist and as a background for the particular problem of grain size estimation discussed in chapter 5.

Chapter 5 presents a program for estimating the average grain size of metals based on the three circle (Abrams) procedure defined in ASTM Designation: E 112-74 "Standard Methods for Estimating the Average Grain Size of Metals". The program makes use of multiple edge mask filters to determine grain boundaries crossed by each circular test pattern. Grain crossings are determined by comparing the relative response of three different mask sizes. This makes the program usable over a broader range of image magnifications and grain sizes and, at the same time, makes it less susceptible to false indications due to noise. The program is demonstrated using a sample turbine airfoil. An apparent change in grain size from one region of the image to another may be due to a change in grain size or to a change in the view angle. In order to determine actual grain size, it is necessary to explicitly account for the foreshortening effect of the view angle at each image point. This can be done if the surface orientation is known at each image point.

Chapter 6 illustrates the experimental determination of a reflectance map for the gray iron used in the green sand mold casting of shuttle eyes for textile looms. An improvised photo-goniometer is used to make the required reflectance measurements. An improvised coordinate gauging machine is used to determine object relief. The results are combined to produce a synthesized image of the shuttle eye.

Chapter 7 presents some concluding remarks concerning the research.

Appendix A summarizes mathematical details that were ommitted in the body of the report.

Appendix B presents a glossary of typical casting defects.

2. INTERPRETING IMAGE INTENSITY

This chapter develops a framework for interpreting the intensity values recorded in an image. The goal is to relate image intensity to topographic properties of the object being imaged. Three such properties are considered:

- Range
- Surface Orientation
- Surface Curvature

These properties are the basis for a viewer-centered representation of object shape. The relation between these properties and image intensity is determined by the surface photometry of the objects in view, by the nature of the incident illumination and by the object surface, light source and viewer geometry. This relation is made explicit by examining the basic image-forming equation:

<image intensity> = <incident illumination> x <surface reflectance> (2.1)

Analysis of this equation is used to show how image intensity relates to surface orientation and how changes in image intensity relate to surface curvature.

Image analysis is formulated as the problem of determining, at each image point, the range, the surface orientation and the surface curvature at the corresponding object point. *Range* is the distance of the object point from the viewer. It is a one-parameter function at each image point. *Surface orientation* is the direction of a viewer-facing surface normal at the object point. It is a two-parameter function at each image point. *Surface curvature* is the two principal curvatures, and associated directions, at the object point. It is a three-parameter function at each

image point. Taken together, range, surface orientation and surface curvature define the *object relief* at each image point.

Analysis of equation (2.1) is facilitated using the notions of gradient space and a reflectance map. Gradient space is a convenient way of representing surface orientation. A reflectance map is a convenient way of representing image intensity as a function of surface orientation. With these tools, the basic image-forming equation takes on a simple form. Constraints imposed by the light source, object surface and viewer geometry are made explicit. The reflectance map is the basis for photometric methods to determine the range, surface orientation and surface curvature at object points corresponding to selected image points.

A particular relaxation algorithm is presented for determining surface orientation from a single view by propagation of mutual constraint on possible orientations at selected image points consistent with general hypotheses about object shape. Constraints on surface curvature are translated into geometric constraints on possible surface orientations that can correspond to a given image point. This algorithm serves to illustrate that the local interpretation of image intensity depends upon local assumptions about surface curvature.

2.1 A VIEWER-CENTERED REPRESENTATION FOR SHAPE

In image analysis, representations used depend more upon what it is possible to compute from the intensity values in an image than upon what is ultimately desirable. In subsequent stages of analysis, representations can be made more sensitive to the needs of the specific application.

A viewer-centered representation is appropriate for the initial determination of the shape of objects in an image. The representation chosen distinguishes the following three classes of image feature:

1. Points of range discontinuity (occlusion boundaries).
2. Points of surface-orientation discontinuity (convex/concave edges).
3. Sections of smooth surface.

[Marr 77b] calls this representation the 2-1/2-D sketch and argues for its usefulness as an intermediate stage in a vision system for determining an object-centered description of shape. It is useful to explore how this viewer-centered 2-1/2-D sketch can be determined directly from the intensity values recorded in an image. The particular question addressed in this chapter is how the range, surface orientation and surface curvature can be determined over sections of smooth surface.

2.2 DETERMINING OBJECT RELIEF

Before developing methods based upon image analysis, it is worthwhile to review other possible techniques for determining object relief at a distance. Figure 2-1 summarizes the possibilities.

Roughly speaking, methods for determining object relief can be divided into two categories. First, there are methods which attempt to measure range directly. Such methods can be based on time-of-flight measurements of a signal reflected back from the object surface. One design uses a pulsed laser to measure time-of-flight. This approach has proven practical for the long distance ranging required in space exploration [Johnston 73]. It has also been applied to more terrestrial endeavors [Caulfield 76]. A more common design for short distance ranging measures time-of-flight as a

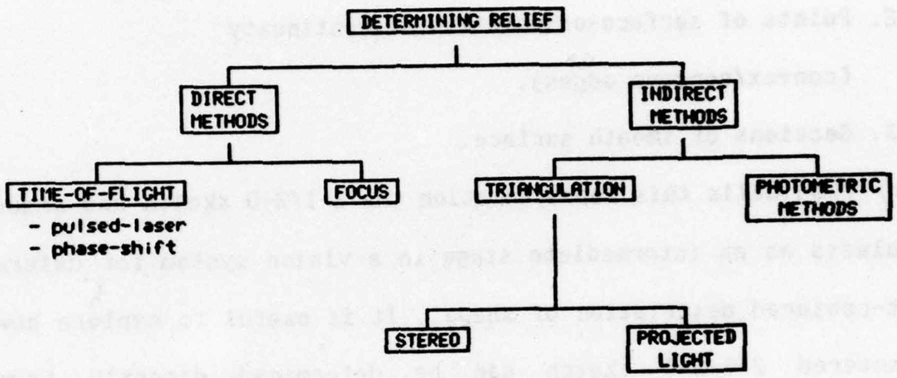


Figure 2-1 Techniques for determining object relief.

phase shift between the emitted and returning signal. The CW LIDAR Pointing System developed at the C. S. Draper Laboratory measures time-of-flight this way using a noncoherent LED centered at 910 nanometers and amplitude modulated at 50 MHz. [DRAPER 73]. This system is reported accurate to ±1 millimeter at a total distance of about 3 meters, approximately ±1 part in 3000. A similar system has been developed at Stanford Research Institute and is reported in [Nitzan *et al* 77]. This last reference includes a useful discussion of the inherent design trade-offs associated with such direct ranging devices. Basically, accuracy is achieved by maintaining a favorable signal to noise ratio. The signal to noise ratio can be improved either by increasing the power of the source or increasing the time constant used to sample each point.

Focussing can also be thought of as a direct ranging technique. Focus in an optical system depends on the range of the objects being imaged. Techniques for the automatic focussing of imaging systems have been explored [Horn 68]. In an early experimental system [Winston 73], automatic focussing was able to locate objects in a scene to an accuracy of about ±1 inch at a total distance of about 6 feet, approximately ±1 part in 100. The use of focussing as a ranging technique depends on the presence of sharp features, such as edges, in the scene. It is not suited for ranging over objects with uniform surface cover and smoothly varying topography.

Second, there are methods which determine range from indirect measurements. These techniques can be further subdivided into triangulation methods and photometric methods. One kind of triangulation method uses a stereo pair of images obtained from two known positions. Any point in an image determines a ray in space. The range at a point in an image is determined by computing the intersection of this ray and the ray

determined by the corresponding point in the other image. But, in order to determine the second ray, it is first necessary to find the corresponding point in the other image. Thus, stereo ranging suffers from the same qualitative limitations as does the technique of automatic focussing. Solving the correspondence problem depends on the presence of features in one image that can be matched to features in the other. It is unsuitable for ranging over objects with uniform surface cover and smoothly varying topography.

Another kind of triangulation method avoids the correspondence problem by using specially controlled illumination. The idea is to project thin sheets of light sequentially on the scene and record the image resulting from each sheet. The range at a point in an image is determined by computing the intersection of the ray associated with the image point and the sheet of light which illuminated the point.

It is also possible to use specially controlled illumination to determine surface orientation. [Will & Pennington 71] describe a method which uses special illumination to code surface orientation as the modulation on a spatial frequency carrier grid. With this method, it is possible to extract the surface orientation of planar regions in a scene by linear frequency domain filtering.

Finally, as indicated above, photometric methods are methods which indirectly determine range from analysis of the image-forming equation (2.1). A photometric approach to range determination was first proposed in [van Diggelen 51] and subsequently applied to lunar images obtained from Ranger spacecraft [Rindfleisch 66]. The particular reflectance properties of the material of the maria of the moon cause a simplification in the mathematics required to determine range from intensity. [Horn 75]

generalized the photometric approach to account for surfaces with arbitrary isotropic reflectance properties. Photometric methods determine surface orientation from image intensity. Once the surface orientation at each object point is determined, range values are obtained by starting at a known point and integrating surface orientation over sections of smooth surface. Photometric methods are complementary to stereo ranging and focus ranging because photometric methods are most suited to objects with uniform surface cover and smoothly varying topography.

Regardless of how object relief is determined, it is important to realize that object relief, in the form of a 2-1/2-D sketch, corresponds to an image with effects due to incident illumination and surface photometry explicitly factored out. Thus, exploring how to determine object relief directly from the intensity values recorded in an image is important to image analysis because it addresses the issue of how to account for the effects of incident illumination and surface photometry.

Note, however, that two stumbling blocks remain. First, determining object relief does nothing to overcome the problem of the quantity of data present in an image. Indeed, if object relief is determined at each image point, the quantity of data and associated computational load increases significantly. But, subsequent scene analysis has a more explicit representation to deal with and thus can do a better job [Nevatia & Binford 73] [Marr 77b]. Second, while knowledge of object relief solves part of the problem associated with image projection (eg. depth discontinuities can be used to separate objects in an image), it nevertheless corresponds to a viewer-centered representation of shape. Additional assumptions are required in order to recover the three-dimensional structure lost in image projection.

2.3 GRADIENT SPACE AND THE REFLECTANCE MAP

In order to understand the correspondence between image intensity and object relief, it is necessary to relate the geometry of image projection to the radiometry of image formation. But, understanding the radiometry of image formation requires a model of the way surfaces reflect light.

The fraction of light reflected by a surface in a given direction depends upon the optical properties of the object material, the surface microstructure and the spatial and spectral distribution and state of polarization of the incident light. A key photometric observation is:

> *No matter how complex the distribution of incident illumination, for most surfaces, the fraction of the incident light reflected in a particular direction depends only on the surface orientation.*

This section introduces the tools required to take advantage of this observation.

The reflectance characteristics of an object material can be represented as a function $\phi(i,e,g)$ of the three angles *i*, *e* and *g* defined in figure 2-2. These angles are called, respectively, the *incident*, *emergent* and *phase* angles. In this work, the emergent angle e will also be referred to as the *view angle*. The angles i, e and g are defined relative to the object surface. $\phi(i,e,g)$ determines the ratio of radiance to irradiance measured per unit surface area, per unit solid angle, in the direction of the viewer. The reflectance function defined here is related to the bi-directional reflectance-distribution function defined by the National Bureau of Standards [Nicodemus *et al* 77].

Image-forming systems perform a *perspective projection* as illustrated in figure 2-3(a). If the size of the objects in view is small compared to the viewing distance, then the perspective projection can be approximated

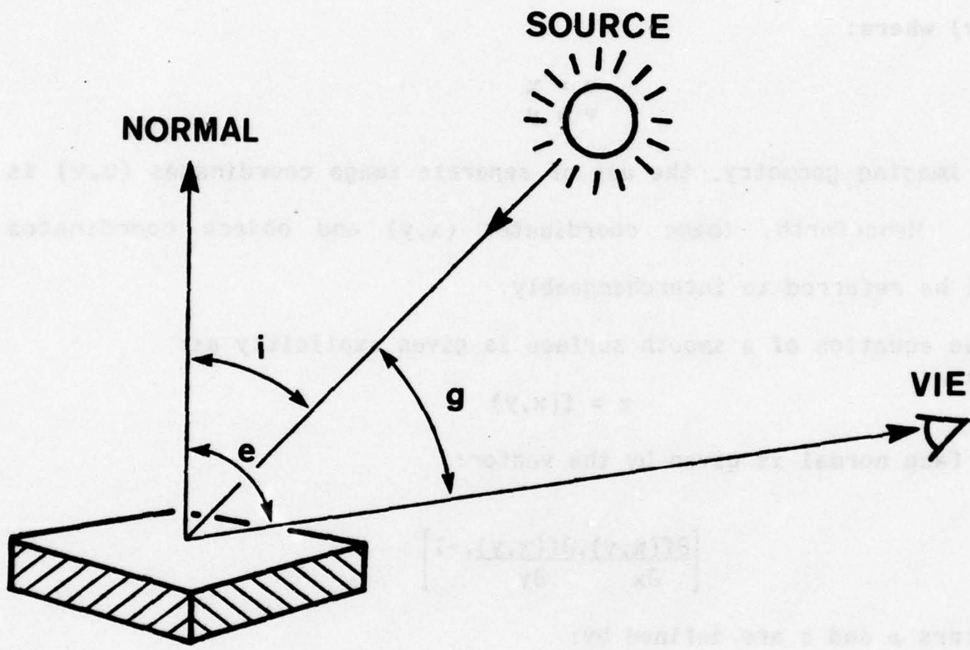


Figure 2-2 Defining the three angles i, e and g. The incident angle i is the angle between the incident ray and the surface normal. The view angle e is the angle between the emergent ray and the surface normal. The phase angle g is the angle between the incident and emergent rays.

as an *orthographic projection* as illustrated in figure 2-3(b).

Consider an image-forming system that performs an orthographic projection. To standardize the imaging geometry, it is convenient to align the viewing direction with the negative z-axis. Also, assume appropriate scaling of the image plane so that object point (x,y,z) maps onto image point (u,v) where:

$$u = x$$
$$v = y$$

With this imaging geometry, the use of separate image coordinates (u,v) is redundant. Henceforth, image coordinates (x,y) and object coordinates (x,y) will be referred to interchangeably.

If the equation of a smooth surface is given explicitly as:

$$z = f(x,y)$$

then a surface normal is given by the vector:

$$\left[\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y}, -1\right]$$

If parameters *p* and *q* are defined by:

$$p = \frac{\partial f(x,y)}{\partial x}$$

$$q = \frac{\partial f(x,y)}{\partial y}$$

then the surface normal becomes [p,q,-1]. The quantity (p,q) will be called the *gradient*, and *gradient space* is the two-dimensional space of all such points (p,q). With the viewer looking along the z-axis, [p,q,-1] defines a viewer-facing surface normal.

Gradient space is a convenient way to represent surface orientation. It has been used in scene analysis [Mackworth 73]. In image analysis, it is used to relate the geometry of image projection to the radiometry of

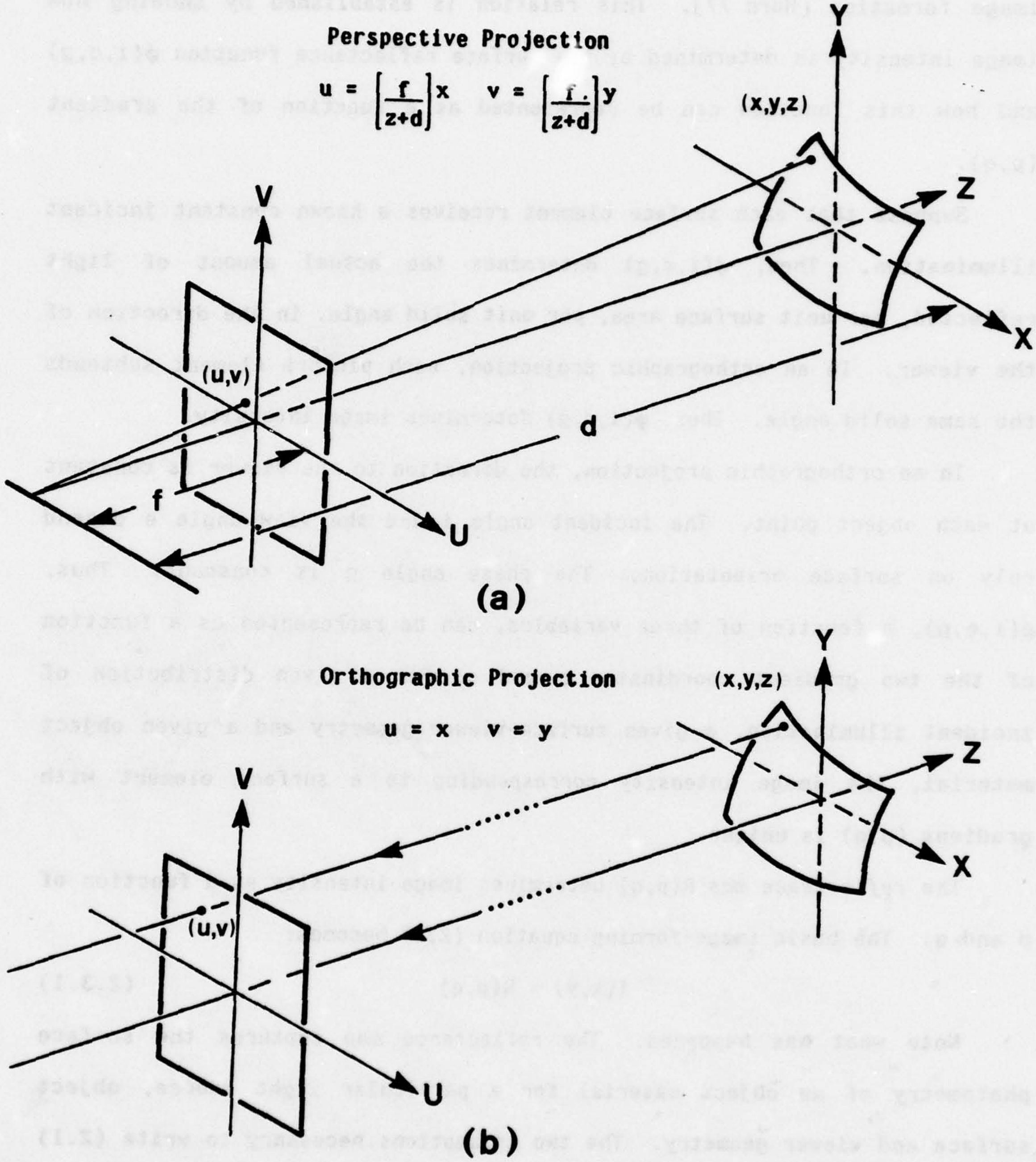


Figure 2-3 Characterizing image projection. To avoid image inversion, it is convenient to think of the image plane as in front of the lens rather than behind it. Figure 2-3(a) illustrates a perspective projection, where d is the distance from the object coordinate origin to the center of the lens and f is the focal length. For objects that are small relative to the viewing distance, the image projection becomes orthographic. In an orthographic projection, illustrated in figure 2-3(b), the focal length f is infinite so that all rays from object to image plane are parallel.

image formation [Horn 77]. This relation is established by showing how image intensity is determined by the surface reflectance function $\phi(i,e,g)$ and how this function can be represented as a function of the gradient (p,q).

Suppose that each surface element receives a known constant incident illumination. Then, $\phi(i,e,g)$ determines the actual amount of light reflected, per unit surface area, per unit solid angle, in the direction of the viewer. In an orthographic projection, each picture element subtends the same solid angle. Thus, $\phi(i,e,g)$ determines image intensity.

In an orthographic projection, the direction to the viewer is constant at each object point. The incident angle i and the view angle e depend only on surface orientation. The phase angle g is constant. Thus, *$\phi(i,e,g)$, a function of three variables*, can be represented as a function of the two gradient coordinates p and q. For a given distribution of incident illumination, a given surface-viewer geometry and a given object material, the image intensity corresponding to a surface element with gradient (p,q) is unique.

The *reflectance map* $R(p,q)$ determines image intensity as a function of p and q. The basic image-forming equation (2.1) becomes:

$$I(x,y) = R(p,q) \tag{2.3.1}$$

Note what has happened. The reflectance map captures the surface photometry of an object material for a particular light source, object surface and viewer geometry. The two assumptions necessary to write (2.1) as (2.3.1) are:

1. The incident illumination is constant at each surface element.
2. The image projection is orthographic.

Reflectance maps can be determined empirically, derived from phenomenological models of surface reflectance or derived from analytic models of surface microstructure. The reflectance map is independent of the shape of the objects being viewed. It represents explicit knowledge of intensities that can be recorded from objects made of a given material and viewed under a particular light source and viewer geometry. A reflectance map is not an image.

Equation (2.3.1) is the basic equation used to relate image intensity to the geometry of the image-forming process. It is one equation in the two unknowns p and q. Thus, the problem of determining surface orientation from intensity becomes the problem of finding the point in gradient space (p,q) corresponding to the image point (x,y).

The simplest case for incident illumination is that of a single distant point source. Suppose such a source exists and that object space vector $[p_s, q_s, -1]$ points in its direction. That is, the source is located in gradient space at point (p_s, q_s). With the imaging geometry of figure 2-3(b), object space vector [0,0,-1] points in the direction of the viewer. That is, the viewer is located in gradient space at point (0,0). Recall that object space vector [p,q,-1] defines a viewer-facing normal at surface point (x,y,z). That is, (p,q) is the gradient point corresponding to surface point (x,y,z).

Then, writing the cosine of the angle between two vectors as the dot product divided by the product of magnitudes, expressions for cos(i), cos(e) and cos(g) become:

$$\cos(i) = \frac{1 + pp_s + qq_s}{\sqrt{1 + p_s^2 + q_s^2}\ \sqrt{1 + p^2 + q^2}}$$

$$\cos(e) = \frac{1}{\sqrt{1 + p^2 + q^2}}$$

$$\cos(g) = \frac{1}{\sqrt{1 + p_s^2 + q_s^2}}$$

Using the above expressions, the transformation of an arbitrary surface reflectance function $\phi(i,e,g)$ into a reflectance map $R(p,q)$ is straightforward. One simple, idealized model of surface reflectance is given by:

$$\phi(i,e,g) = \rho \cos(i)$$

This reflectance function corresponds to the phenomenological model of a perfectly diffuse (lambertian) surface which appears equally bright from all viewing directions. Here, ρ is a reflectance factor and the cosine of the incident angle accounts for the foreshortening of the *surface as seen from the source*. The corresponding reflectance map is *given* by:

$$R(p,q) = \frac{\rho(1 + pp_s + qq_s)}{\sqrt{1 + p_s^2 + q_s^2}\ \sqrt{1 + p^2 + q^2}}$$

It is convenient to present $R(p,q)$ as a series of "iso-brightness" contours in gradient space. Figure 2-4 shows the reflectance map $R(p,q) = \rho \cos(i)$ drawn as a series of contours, for the case $p_s = 0.7$, $q_s = 0.3$ and $\rho = 1$.

Specifying a single distant point source is not a *fundamental restriction*. *Nonpoint* sources can be modeled as the superposition of single point sources. The reflectance map does, however, assume equal *illumination* at all *surface elements*. The reflectance map does not account for the fact that certain surface elements can be shadowed with respect to one or more of the sources nor for the fact that, for nonconvex objects,

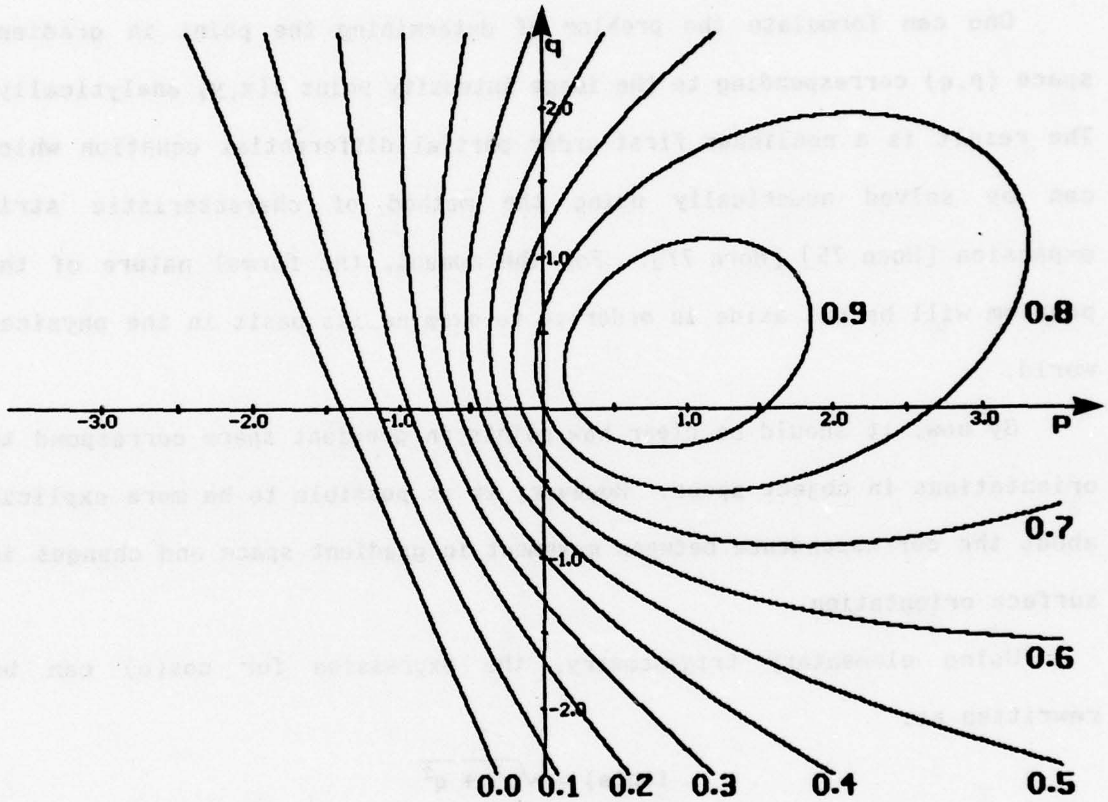


Figure 2-4 The reflectance map for a lambertian surface illuminated by a single distant point source at gradient $p_s = 0.7$ and $q_s = 0.3$ with $\rho = 1$. The reflectance map is plotted as a series of contours spaced 0.1 units apart.

certain surface elements can receive secondary illumination due to light reflected from other sections of surface (i.e., mutual illumination).

2.4 RE-EXAMINING PHYSICAL CONSTRAINTS

One can formulate the problem of determining the point in gradient space (p,q) corresponding to the image intensity point I(x,y) analytically. The result is a nonlinear first-order partial differential equation which can be solved numerically using the method of characteristic strip expansion [Horn 75] [Horn 77]. For the moment, the formal nature of the problem will be put aside in order to re-examine its basis in the physical world.

By now, it should be clear how points in gradient space correspond to *orientations in object space.* However, it is possible to be more explicit about the correspondence between movement in gradient space and changes in surface orientation.

Using elementary trigonometry, the expression for cos(e) can be rewritten as:

$$\tan(e) = \sqrt{p^2 + q^2}$$

The inclination of the surface with respect to the viewing direction varies monotonically with distance from the origin in gradient space. Specifically, the distance from the origin is the tangent of the angle between the surface normal and the viewing direction. As the gradient (p,q) moves away from the origin, the inclination of the surface with respect to the viewer increases.

The view angle e characterizes one of the two degrees of freedom associated with an arbitrary orientation in space. The locus of points in object space having a constant view angle e defines a right circular cone

with axis along the viewing direction. The angular position $\tan^{-1}(q/p)$ of each gradient (p,q) on the circle $p^2 + q^2 = \tan^2(e)$ determines the direction of steepest descent in image space along this cone. In general, the angular position of a point (p,q) in gradient space corresponds to the direction of steepest descent in image space along the object surface. Rotating object space about the viewing direction induces an equal rotation in gradient space.

Two physical constraints can now be identified:

1. A given point on a physical surface has a unique orientation in space.
2. Matter is cohesive. It is separated into objects. The surfaces of objects are generally smooth compared with their distance from the viewer.

These are essentially the same two constraints [Marr & Poggio 76] use as a basis for their method to compute stereo disparity. As in their paper, the problem is to translate the above two physical constraints into rules for how points in an image can be matched to points in gradient space.

In their most general form, these rules can be expressed as:

1. UNIQUENESS:
   Each image point may be assigned to at most one location in gradient space.
2. CONTINUITY:
   Surfaces vary smoothly almost everywhere. Only a small fraction of the area of an image is composed of boundaries that correspond to discontinuities of range or surface orientation.

The task ahead is to demonstrate that these rules can be explicitly embedded in a computation. The result is an algorithm which attempts to achieve a global correspondence between image points and points in gradient space via propagation of local constraints.

Such methods have been called cooperative algorithms [Marr & Poggio 76] or relaxation labelling [Rosenfeld, Hummel & Zucker 76]. Although the implementation has some intrinsic interest, what is important here is to understand the physical basis for the local constraints used and to get the flavor of how these local constraints can propagate back and forth to constrain globally possible matches between image points and points in gradient space.

2.5 SPECIFYING LOCAL CONSTRAINT

The basic image-forming equation

$$I(x,y) = R(p,q)$$

is one equation in the two unknowns p and q. With this equation alone, the gradient corresponding to a particular image point is constrained to lie on a one-parameter family of "iso-brightness" contours in gradient space. The goal is to apply further constraint in order to assign a unique location in gradient space to each image point.

The essential physical constraint to be exploited is that surfaces vary smoothly almost everywhere. This surface smoothness assumption is translated into monotonicity rules on changes to view angle and changes to direction of steepest descent permitted between closely spaced image points.

One can illustrate how physical constraint adds additional constraint to the possible gradient space solutions to the basic equation $I(x,y) = R(p,q)$. Suppose, two closely spaced image points (x_1,y_1) and (x_2,y_2) are hypothesized to correspond to object points on the same section of smooth surface. Further, suppose that the view angle increases in going from (x_1,y_1) to (x_2,y_2) and that the angular position, corresponding to the

direction of steepest descent, decreases in going from (x_1,y_1) to (x_2,y_2). Let (p_1,q_1) and (p_2,q_2) be the gradient locations corresponding to (x_1,y_1) and (x_2,y_2).

Suppose, further, that $I(x_1,y_1) = \alpha_1$ and $I(x_2,y_2) = \alpha_2$. Let C_1 and C_2 be, respectively, the contours in gradient space given by $R(p,q) = \alpha_1$ and $R(p,q) = \alpha_2$ as shown in figure 2-5. Then, (p_1,q_1) lies on C_1 and (p_2,q_2) lies on C_2.

Figure 2-6 shows both the gradient space circle corresponding to the maximum view angle interpretation of (p_2,q_2) (i.e., the circle passing through the point on C_2 furthest from the origin) and the gradient space line corresponding to the minimum direction of steepest descent interpretation of (p_2,q_2) (i.e., the line passing through the point on C_2 with minimum angular position). Since the view angle is assumed to increase in going from (x_1,y_1) to (x_2,y_2), the contour of permissible (p_1,q_1) can be restricted to include only those gradient points on C_1 lying on or within the circle of the maximum view angle interpretation of (p_2,q_2). Similarly, since angular position is assumed to decrease in going from (x_1,y_1) to (x_2,y_2), the contour of permissible (p_1,q_1) can be restricted to include only those gradient points on or above the line of the minimum direction of steepest descent interpretation of (p_2,q_2). Thus, without any additional constraint on (p_2,q_2), the assumed monotonicity relations between (x_1,y_1) and (x_2,y_2) have been applied to the reflectance map contours to constrain the possible interpretation of (p_1,q_1) to include only those points of C_1 indicated by the solid line of figure 2-6.

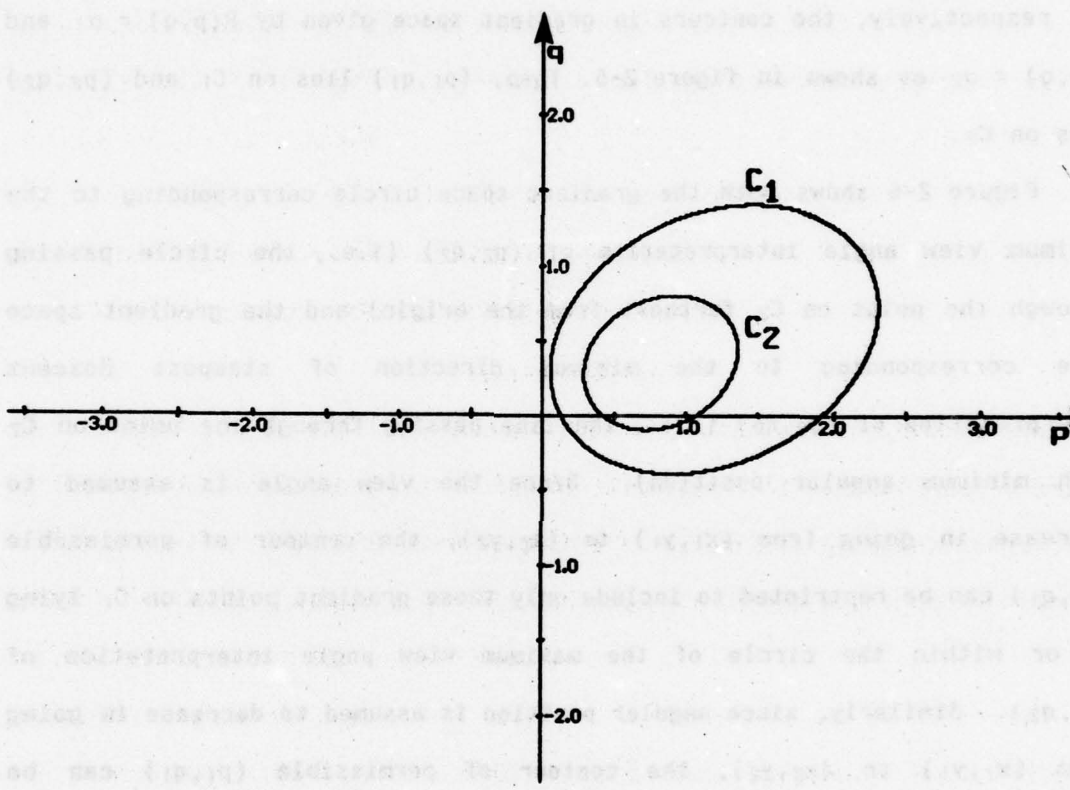


Figure 2-5 The two reflectance map contours C_1 and C_2, corresponding to $R(p,q) = \alpha_1$ and $R(p,q) = \alpha_2$, which determine possible surface orientations at image points (x_1,y_1) and (x_2,y_2).

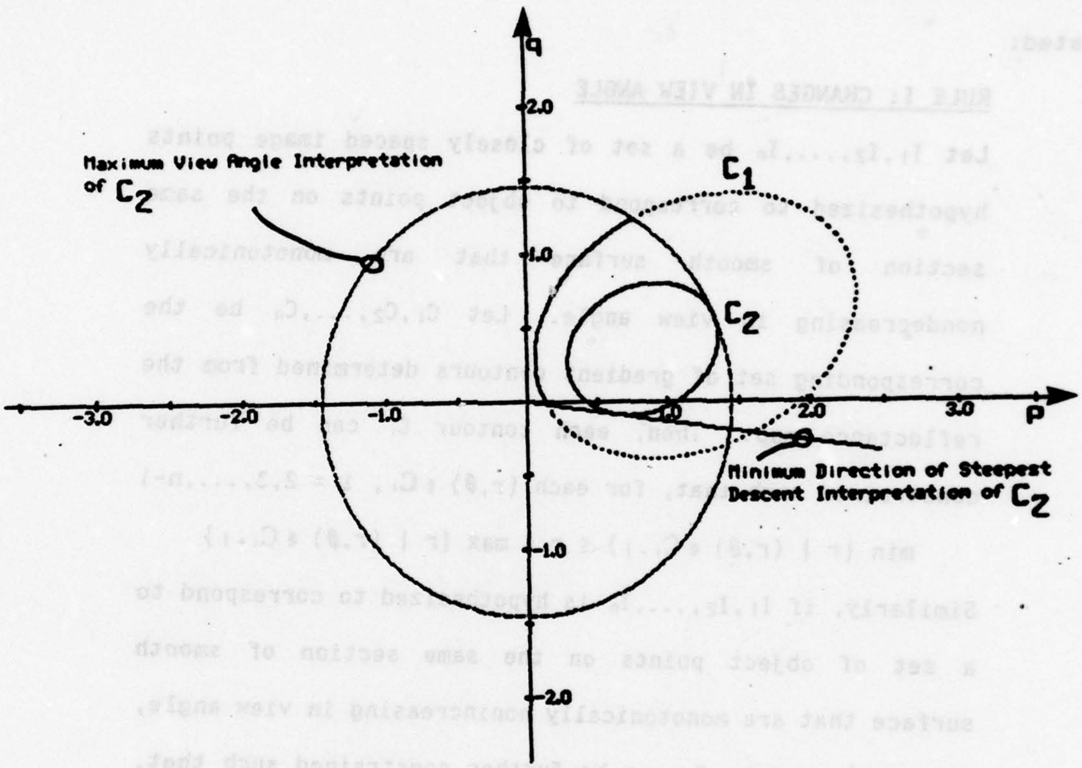


Figure 2-6 The restricted subsection of contour C_1 that is consistent with the hypothesis that the view angle increases and the direction of steepest descent decreases in going from (x_1, y_1) to (x_2, y_2).

Let (r,θ) denote the polar representation of the gradient (p,q). That is:

$$r = \sqrt{p^2 + q^2}$$

$$\theta = \tan^{-1}(q/p)$$

Generalizing from the above illustration, the following two rules are stated:

RULE I: CHANGES IN VIEW ANGLE

Let $I_1, I_2, \ldots, I_n$ be a set of closely spaced image points hypothesized to correspond to object points on the same section of smooth surface that are monotonically nondecreasing in view angle. Let $C_1, C_2, \ldots, C_n$ be the corresponding set of gradient contours determined from the reflectance map. Then, each contour C_i can be further constrained such that, for each $(r,\theta) \in C_i$, $i = 2,3,\ldots,n-1$

$$\min \{r \mid (r,\theta) \in C_{i-1}\} \le r \le \max \{r \mid (r,\theta) \in C_{i+1}\}$$

Similarly, if $I_1, I_2, \ldots, I_n$ is hypothesized to correspond to a set of object points on the same section of smooth surface that are monotonically nonincreasing in view angle, then each contour C_i can be further constrained such that, for each $(r,\theta) \in C_i$, $i = 2,3,\ldots,n-1$

$$\min \{r \mid (r,\theta) \in C_{i+1}\} \le r \le \max \{r \mid (r,\theta) \in C_{i-1}\}$$

RULE II: CHANGES IN DIRECTION OF STEEPEST DESCENT

Let $I_1, I_2, \ldots, I_n$ be a set of closely spaced image points hypothesized to correspond to object points on the same section of smooth surface that are monotonically nondecreasing in direction of steepest descent. As above, let $C_1, C_2, \ldots, C_n$ be the corresponding set of gradient

contours. Then, each contour C_i can be further constrained such that, for each $(r,\theta) \in C_i$, $i = 2,3,\ldots,n-1$

$$\min \{\theta \mid (r,\theta) \in C_{i-1}\} \leq \theta \leq \max \{\theta \mid (r,\theta) \in C_{i+1}\}$$

Similarly, if $I_1, I_2, \ldots, I_n$ is hypothesized to correspond to a set of object points on the same section of smooth surface that are monotonically nonincreasing in direction of steepest descent, then each contour C_i can be further constrained such that, for each $(r,\theta) \in C_i$, $i = 2,3,\ldots,n-1$

$$\min \{\theta \mid (r,\theta) \in C_{i+1}\} \leq \theta \leq \max \{\theta \mid (r,\theta) \in C_{i-1}\}$$

2.6 HYPOTHESIZING MONOTONICITY RELATIONS

It is now time to turn to the question of how to hypothesize monotonicity relations between selected image points. To begin with, consider the worst possible approach. For some small value of n, one might explore all possible orderings, with respect to both view angle and direction of steepest descent, of selected image points $I_1, I_2, \ldots, I_n$. The hope would be that only a small fraction of those orderings would have admissable interpretations (i.e., interpretations that included at least one gradient point for each image point). The constraints imposed by each interpretation would propagate to neighboring sets of selected image points to provide further mutual constraint. Again, the hope would be that propagation of local constraint would converge to a correct global interpretation while "incorrect" propagations would quickly die out.

Consider a second possible approach. Suppose a particular surface interpretation is forced onto the data. Such an interpretation would provide a framework to partially order selected, closely spaced, image points with respect to changes in both view angle and direction of steepest

descent. Instead of allowing all possible orderings to compete, this second approach pursues a particular interpretation. Again, the hope would be that propagation of local constraint would converge to a single global interpretation that represents a simple distortion of the particular interpretation being forced. Here, simple distortion implies any surface that preserves the assumed monotonicity relations concerning changes to view angle and changes to direction of steepest descent.

The method actually implemented corresponds to this second approach. The program has a small set of interpretations it is willing to pursue. Some are quite rigid, others are quite flexible. In the next section, a specific example is presented. For now, a brief analysis is given to show how convexity can be used to hypothesize monotonicity relations between closely spaced image points. In the process, an important theory will be developed. Appendix A contains a more detailed formulation of the mathematical results presented here.

By taking partial derivatives of the basic equation $I(x,y) = R(p,q)$ with respect to x and y two equations are obtained which can be written as the single matrix equation:

$$\begin{bmatrix} I_x \\ I_y \end{bmatrix} = \begin{bmatrix} p_x & q_x \\ p_y & q_y \end{bmatrix} \begin{bmatrix} R_p \\ R_q \end{bmatrix} \tag{2.6.1}$$

(Throughout, subscripts are used to denote partial differentiation). Similarly, the two first-order equations:

$$dp = p_x\,dx + p_y\,dy$$

$$dq = q_x\,dx + q_y\,dy$$

can be written as the single matrix equation:

$$\begin{bmatrix} dp \\ dq \end{bmatrix} = \begin{bmatrix} p_x & p_y \\ q_x & q_y \end{bmatrix} \begin{bmatrix} dx \\ dy \end{bmatrix} \tag{2.6.2}$$

For smooth surfaces, the order of differentiation can be interchanged. Recalling the original definitions of p and q, this means that

$$p_y = q_x$$

One can define a matrix **H** by:

$$\mathbf{H} = \begin{bmatrix} \dfrac{\partial^2 f(x,y)}{\partial x^2} & \dfrac{\partial^2 f(x,y)}{\partial x \partial y} \\ \dfrac{\partial^2 f(x,y)}{\partial x \partial y} & \dfrac{\partial^2 f(x,y)}{\partial y^2} \end{bmatrix}$$

H is the standard Hessian matrix of the function $z = f(x,y)$. Here, **H** is called the *image Hessian matrix* of the surface $z = f(x,y)$. **H** is a viewer-centered representation of surface curvature. In Appendix A.5, the image Hessian matrix **H** is related directly to an object-centered definition of curvature.

Equation (2.6.2) can be written in the form:

$$[dp,dq]^T = \mathbf{H}\,[dx,dy]^T$$

The image Hessian matrix **H** relates movement in the image to the corresponding movement in gradient space.

The particular result to be used in this section is that the definiteness of **H** is related to the convexity/concavity of the object surface $z = f(x,y)$:

> *The object surface $z = f(x,y)$ is convex with respect to the viewer if and only if the corresponding image Hessian matrix* **H** *is positive semidefinite.*

Similarly,

> *The object surface $z = f(x,y)$ is concave with respect to the viewer if and only if the corresponding image Hessian matrix* H *is negative semidefinite.*

Suppose $z = f(x,y)$ is convex. Then, H is positive semidefinite. Multiplying the two matrix equations (2.6.1) and (2.6.2) on the left by $[R_p \; R_q]$ and $[dx \; dy]$ respectively, gives rise to the two inequalities:

$$R_p \, I_x + R_q \, I_y \geq 0 \tag{2.6.3}$$

$$dp \, dx + dq \, dy \geq 0 \tag{2.6.4}$$

Two similar inequalities hold, with the sense of the inequality reversed, if $z = f(x,y)$ is concave. Note, however, that concavity need not be treated as a separate case. Indeed, ignoring shadows and mutual illumination, the classic indentation/protrusion ambiguity has a simple expression in this framework.

> *If $I(x,y)$ is the image corresponding to a concave surface $z = f(x,y)$ illuminated by a single point source at gradient point (p_s, q_s) then $I(x,y)$ is also the image corresponding to the convex surface $z = -f(x,y)$ illuminated by a single point source at gradient point $(-p_s, -q_s)$.*

The first inequality (2.6.3) is an additional constraint on the gradient corresponding to a point on a convex surface $z = f(x,y)$. The normal vector $[R_p, R_q]$ to the contour of constant reflectance at any point (p,q) hypothesized to be a solution to the basic equation $I(x,y) = R(p,q)$ must have a nonnegative component in the direction of the normal vector $[I_x, I_y]$ to the contour of constant intensity at (x,y).

The second inequality (2.6.4) can be viewed as an additional constraint on the possible movement $[dp,dq]$ in gradient space corresponding to a movement $[dx,dy]$ in the image. The vector $[dp,dq]$ must have a nonnegative component in the direction $[dx,dy]$. In Appendix A.2, it is shown how to choose $[dx,dy]$ to guarantee either the sign of the change to

the view angle or the sign of the change to the direction of steepest descent.

2.7 ACHIEVING GLOBAL CONSTRAINT

Regardless of what mechanism is used to hypothesize monotonicity relations between points in image space, it is still necessary to embed that mechanism in a computation to achieve global constraint. The implementation approach taken here is somewhat ad hoc. The program selects nine points in a 3 x 3 square pattern as its basic set of closely spaced image points. This pattern serves as the set $I_1, I_2, \ldots, I_n$ for applying the local constraint criteria (Rules I and II above). First, however, the set $I_1, I_2, \ldots, I_n$ is passed to the chosen hypothesizing routines to be partially ordered with respect to view angle and direction of steepest descent. The reflectance map $R(p,q)$ is then used to determine the initial contour of possible gradient space locations for each point I_i. Rules I and II are iteratively applied to these contours until no further mutual constraint is provided.

The above describes the basic application of local constraint to each 3 x 3 square pattern. The selection of successive 3 x 3 patterns is allowed to overlap. Thus, each image point I_i will eventually belong to nine 3 x 3 patterns. Each time a particular image point I_i is further constrained by the application of local constraint to a pattern of which it is a member, each of its eight other patterns is marked for reconsideration. Before moving on to a previously unconsidered pattern local constraint is applied iteratively to each marked pattern, with additional marking added as required, until no marked patterns remain to be reconsidered. Each time an image point I_i is considered, any additional

constraint on the gradient space contour of possible solutions to the basic equation $I(x,y) = R(p,q)$ propagates through this local filtering mechanism to all other image points under consideration.

The next issue to arise is the question of how to terminate the growth of patterns. Currently, the program terminates on one of two conditions:

1. One or more of the image points under consideration has no admissable gradient space interpretation. This means the algorithm can no longer assign surface orientations to image points consistent with the hypotheses about surface curvature used to generate the partial orderings with respect to changes to view angle and changes to direction of steepest descent. In this case, the "forced" interpretation is deemed to have failed.
2. Any new patterns will cross either a contour of range discontinuity (OCCLUSION boundary) or a contour of discontinuity of surface normal (CONVEX or CONCAVE edge). OCCLUSION boundaries are detected by noting when boundary points to existing patterns have a view angle greater than some preassigned value. (That is, e is approaching $\pi/2$). CONVEX or CONCAVE edges, on the other hand, cannot be detected while patterns are being expanded. Unless external knowledge is provided about the existence of such edges, patterns will propagate across discontinuities in surface normal. Depending on the nature of the discontinuity, it may lead to failure due to condition 1 above or it may never be noticed.

2.8 AN ILLUSTRATIVE EXAMPLE

Consider the simple example of a lambertian sphere illuminated by a single distant light source. The intensity space to gradient space correspondence will be derived analytically and used to generate a synthetic image. The algorithm will be applied to the synthesized image. The results will be used as a basis for judging the performance of the algorithm.

The first task is to determine the reflectance map. As above, the reflectance map for a lambertian surface illuminated from gradient point (p_s, q_s) is given by:

$$R(p,q) = \frac{\rho(1 + pp_s + qq_s)}{\sqrt{1 + p_s^2 + q_s^2}\ \sqrt{1 + p^2 + q^2}}$$

For the example, $p_s = 0.7$, $q_s = 0.3$ and $\rho = 1$. This reflectance map has been illustrated in figure 2-4.

The second task is to determine the surface orientation at each image point. For the example, this result is easy to derive. Let the sphere be centered at the object space origin and have radius r. Then, the equation of the sphere is given implicitly by:

$$x^2 + y^2 + z^2 = r^2$$

Elementary calculus will verify that the vector [x,y,z] defines an outward surface normal at each object point (x,y,z). The corresponding gradient is obtained by rewriting this normal as [-x/z,-y/z,-1]. For each (x,y), there are two possible z values. With the viewer looking along the positive z-axis, however, the hemisphere in view corresponds to negative values of z. For this imaging geometry, the explicit equation of the sphere is given by:

$$z = -\sqrt{r^2 - x^2 - y^2}$$

The synthetic image intensity $I(x,y)$ is equal to the value of the reflectance map at the corresponding (p,q). The equation for image intensity is:

$$I(x,y) = \begin{cases} 0 & \text{if } x^2 + y^2 > r^2 \\ \max\{0, R(-x/z,-y/z)\} & \text{otherwise} \end{cases}$$

where $z = -\sqrt{r^2 - x^2 - y^2}$

Finally, the mechanism by which monotonicity relations were hypothesized for this example must be specified. One additional assumption was used. The algorithm assumes that it knows at least one image point corresponding to an object point oriented directly facing the viewer. Now, let such a point define a *pseudo-origin* in image space. Since the view angle is assumed to be zero at the pseudo-origin, the only possible interpretation is that, in a particular direction, the view angle is locally nondecreasing with increasing image distance from the pseudo-origin. In general, one cannot hope to assert any local monotonicity relation on direction of steepest descent based on angular position about the pseudo-origin. If, however, the surface is known to be convex, the direction of steepest descent is locally nondecreasing with increasing angular position about the pseudo-origin. For the example, the set of image points $I_1, I_2, \ldots, I_n$ was ordered in view angle according to their distance in image space from the pseudo-origin and ordered in direction of steepest descent according to their angular position about the pseudo-origin. Applied together, these hypotheses are equivalent to the strong assumption that the surface in question is a convex solid of revolution, with the axis of revolution along the viewing direction.

For the example, r = 60. A 128 x 128 synthesized image was generated corresponding to the values $-64 \le x \le 63$ and $-64 \le y \le 63$. The algorithm was applied to this image using 3 x 3 square patterns sampled at an image spacing of 5 points in both X and Y. The pseudo-origin was defined as x = 0 and y = 0. The results are presented as a series of figures. Figure 2-7 is the synthesized image. Figure 2-8 shows the points in gradient space, determined analytically, corresponding to the sampled image points. Figure 2-9 redraws figure 2-8 but with the gradient points corresponding to the self-shadowed portion of the sphere excluded. All gradient points to the left of the contour R(p,q) = 0 correspond to surface points oriented more than 90° away from the direction of incident illumination.

Figure 2-10 shows the restricted subsection of contour determined by the algorithm for each sampled point of figure 2-7. Finally, figure 2-11 superimposes the correct gradient point on each subsection of contour of figure 2-10 to illustrate how well the algorithm has performed. The crosses mark the correct gradient points, determined analytically, while the corresponding subsection of contour marks how well the algorithm has isolated those points.

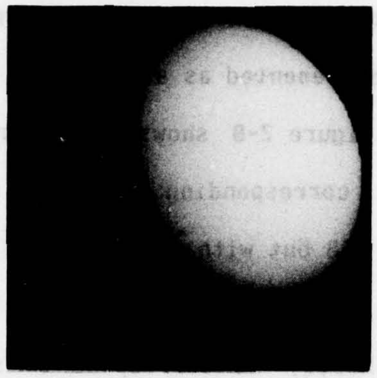


Figure 2-7 The synthesized image of a sphere. The surface reflectance is assumed to be lambertian with a single distant light source at gradient point $p_s = 0.7$ and $q_s = 0.3$.

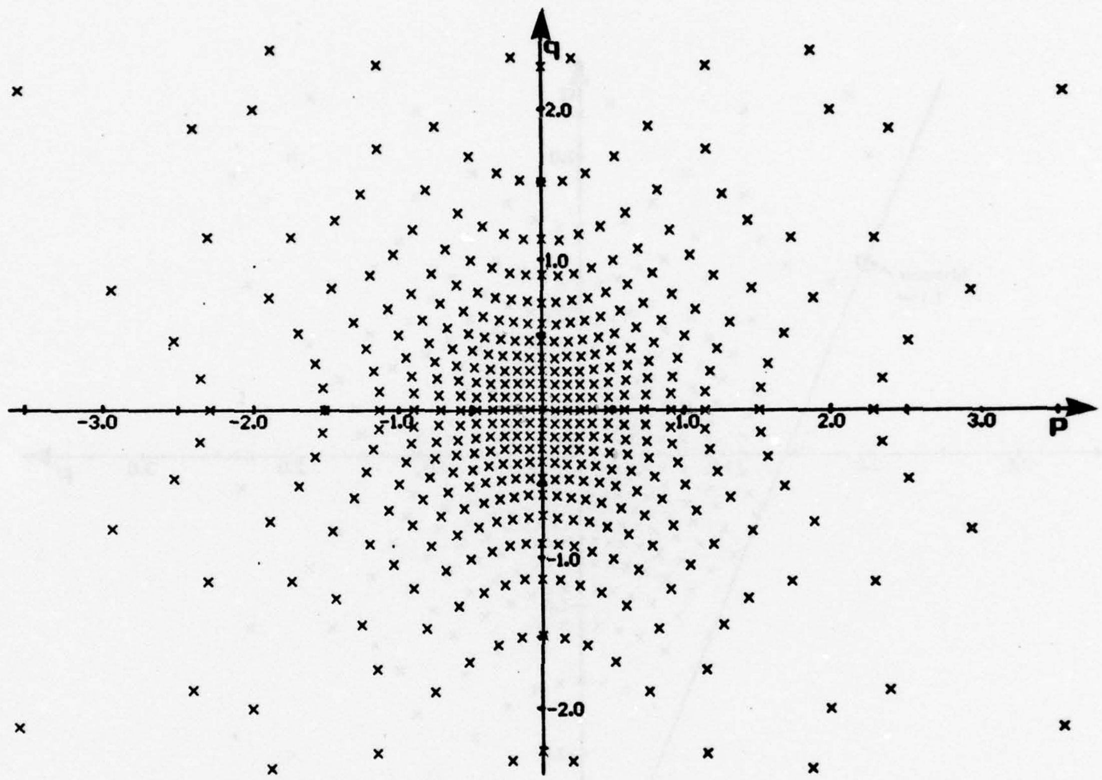


Figure 2-8 The gradient points, determined analytically, corresponding to the sampled image points.

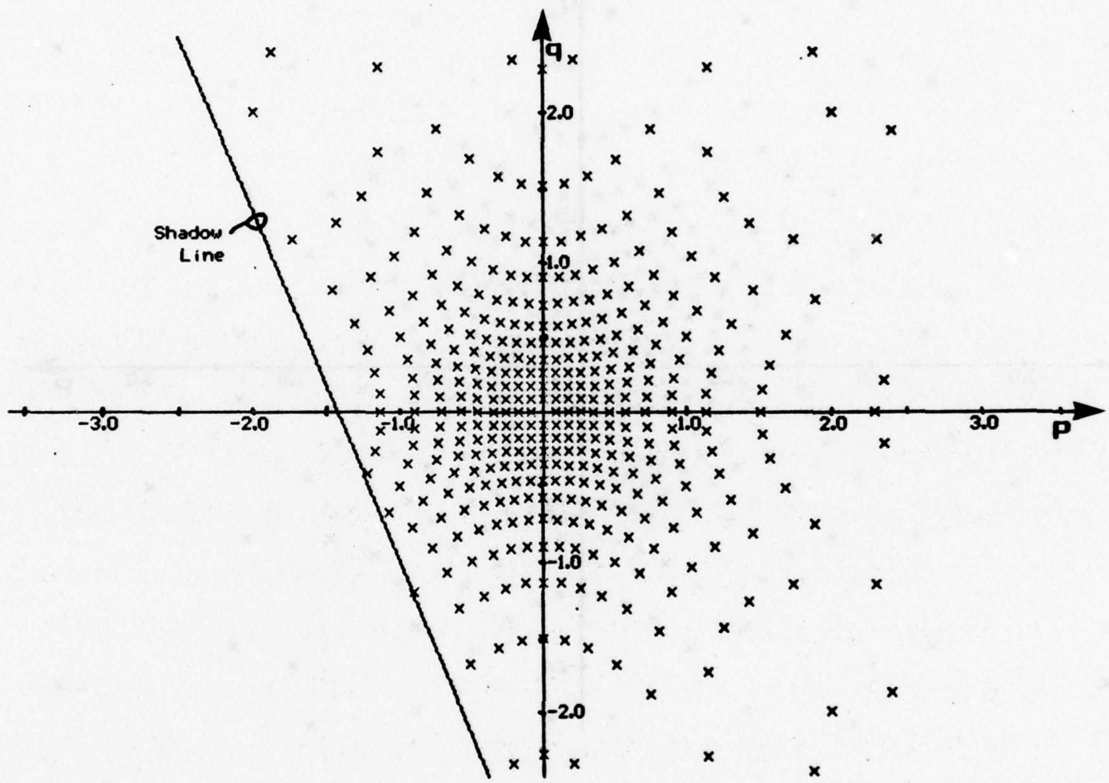


Figure 2-9 The gradient points of figure 2-8 which receive illumination from the light source (i.e., the gradients with $i < \pi/2$).

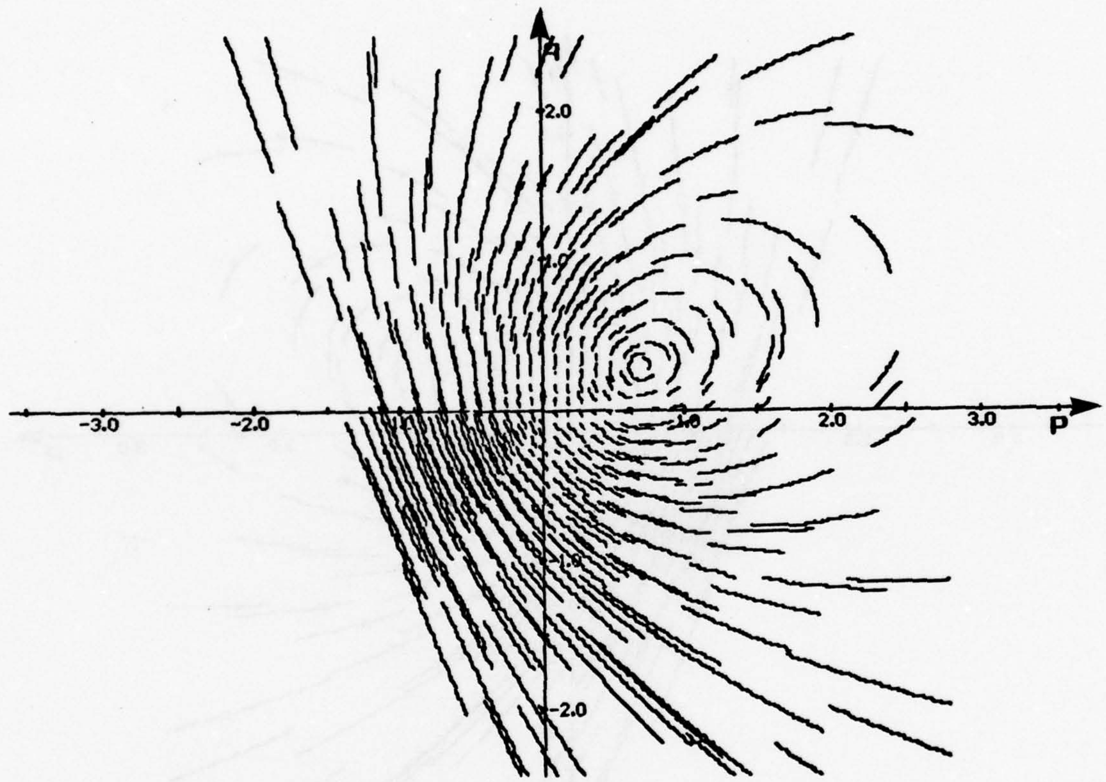


Figure 2-10 The restricted subsection of contour determined for each sampled image point.

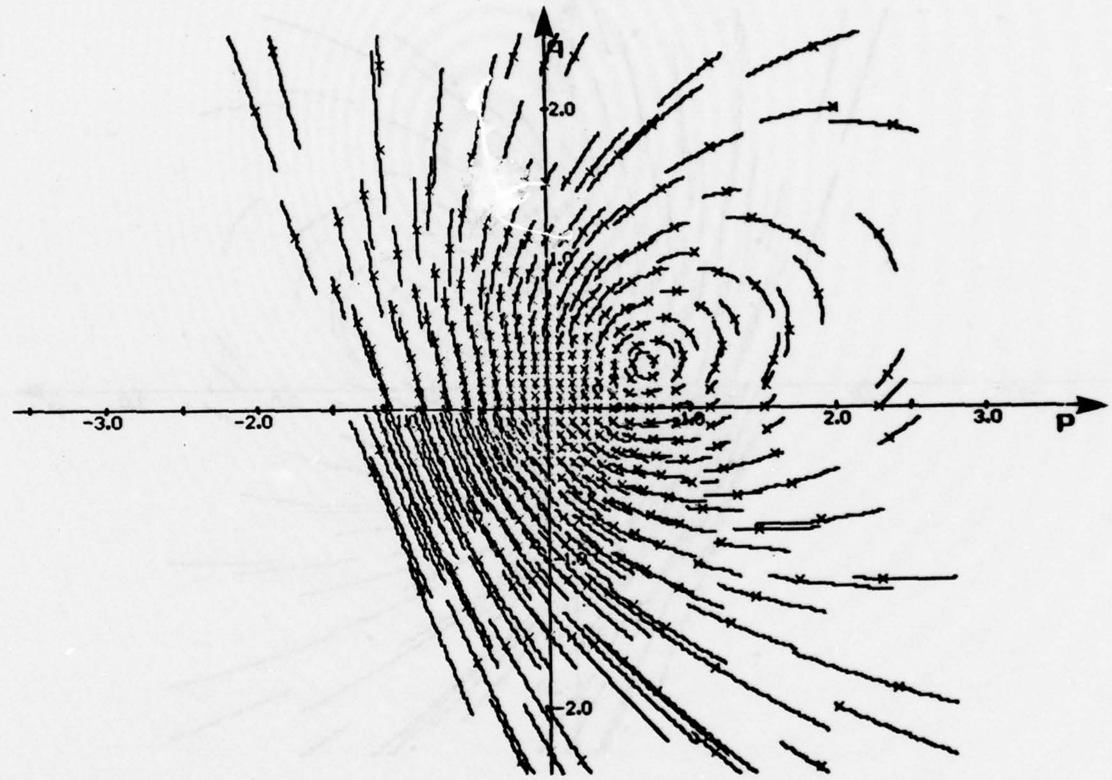


Figure 2-11 The superposition of figure 2-9 and figure 2-10. Crosses mark the exact gradient, determined analytically, while each subsection of contour indicates how well the program has determined that gradient.

2.9 DISCUSSION

The performance of the algorithm can be judged by examining the angular freedom remaining in each subsection of contour C_i. This angular freedom is measured in two ways. First, since each point in gradient space defines a surface orientation, the maximum angle subtended between points remaining in each C_i can be measured. Define the ***angular spread in surface orientation*** at image point I_i as:

$$\max\{\theta \mid \cos(\theta) = \frac{[p_1,q_1,-1].[p_2,q_2,-1]}{|[p_1,q_1,-1]||[p_2,q_2,-1]|} \text{ and } (p_1,q_1),(p_2,q_2)\epsilon C_i\}$$

Second, define the ***angular spread in view angle*** at image point I_i as:

$$\max\{(e_1-e_2) \mid \tan(e_1)=\sqrt{p_1{}^2+q_1{}^2}, \tan(e_2)=\sqrt{p_2{}^2+q_2{}^2} \text{ and } (p_1,q_1),(p_2,q_2)\epsilon C_i\}$$

To tabulate the results, the sampled image points were split into two classes. First, consider all sample points corresponding to object points within 45^o view angle (i.e., lying within gradient space circle $p^2 + q^2 = 1$). Second, consider all sample points corresponding to object points within 60^o view angle (i.e., lying within gradient space circle $p^2 + q^2 = 3$). Table 2-1 summarizes the results:

	Points Within 45° View Angle	Points Within 60° View Angle
<u>Surface Orientation:</u>		
Mean Angular Spread	8.9°	9.9°
Standard Deviation	6.1°	7.0°
Best Case	0.5°	0.5°
Worst Case	36.3°	41.7°
<u>View Angle:</u>		
Mean Angular Spread	4.7°	5.0°
Standard Deviation	2.4°	2.5°
Best Case	0.5°	0.5°
Worst Case	10.9°	11.9°

TABLE 2-1

These measures reflect total angular spread. If a choice algorithm is adopted which selects the "correct" answer to be at the midpoint of the angular spread, this choice is guaranteed to be no more than half the angular spread in error. Thus, the upper right portion of table 2-1 can be interpreted as follows:

> *On the average, by sampling at an image spacing of 5 points in X and Y, the algorithm was able to position image points, corresponding to surface points less than 60° in view angle, to within 5° of their true orientation in space. The standard deviation of this measure over all such points was 3.5° while the worst case point was located to within 21° of its true orientation in space.*

The performance of the algorithm depends critically on three factors: the ability to hypothesize monotonicity relations between selected image points, the nature of constraint propagation and the nature of the reflectance map. Hypothesis based mechanisms are always faced with a chicken and egg dilemma. If one has no a priori constraint on surface curvature, one can say nothing about ordering image points with respect to changes in view angle and changes to direction of steepest descent. On the other hand, the more constraint on surface curvature, the more reliably one

can order image points with respect to changes in view angle and changes to direction of steepest descent. The discussion of convexity gives some indication of how one can use general a priori assumptions about surface curvature to specify monotonicity relations. In the next chapter, situations are explored in which additional curvature constraint can be exploited to simplify image analysis.

In the example above, the hypothesis mechanism used always totally ordered each of the nine pattern points. Still, the algorithm did not determine each gradient exactly. In some sense, it should have. For a solid of revolution, with axis of rotation about the viewing direction, the direction of steepest descent at each image point is given by $\theta = \tan^{-1}(y_p/x_p)$ where (x_p, y_p) is a pseudo-origin as defined above. The algorithm propagates constraint by casting away sections of contour that are inconsistent with the current hypothesis criteria. If the hypothesis criteria are valid, the algorithm will never cast away a correct solution to the equation $I(x,y) = R(p,q)$. On the other hand, this does not imply that each point remaining on a subsection of contour is a possible solution to the equation $I(x,y) = R(p,q)$ consistent with the current hypothesis criteria. The algorithm, as implemented, does no point by point analysis. Any case analysis is a risky business in a numerical algorithm which is, in principle, dealing with a continuous rather than discrete domain. The ability to achieve a tight global constraint is, unfortunately, related to the ability to achieve a tight local constraint. This, in turn, is related not just to the success of the hypothesis mechanism but also to the nature of the reflectance map.

In the example, surface orientation was most poorly constrained in the third and fourth quadrants of gradient space. In the third quadrant, the reflectance map contours locally approximate sections of circles centered at gradient space origin. Here, view angle is tightly constrained but the shape of the reflectance map contours provides little constraint on direction of steepest descent. In the fourth quadrant, the reflectance map contours locally approximate radial lines emanating from the gradient space origin. Here, the direction of steepest descent is tightly constrained but the shape of the reflectance map contours provides little constraint on view angle.

The nature of the reflectance map is determined by two factors: the surface photometry of the object being viewed and the light source, object surface and viewer geometry. In general, one would not expect to have much control over the surface photometry of the objects being viewed. In principle, however, one is free to vary the light source position to achieve optimal results with the algorithm. This idea is the basis for the method of photometric stereo which determines surface orientation using multiple images taken under the same object surface and viewer geometry but with different light source positions (see Chapter 3.4).

The notions of local surface orientation, gradient space and the reflectance map are important tools for image analysis. The beauty of the gradient space formulation is that it allows physical constraints on the object surface to be expressed as simple geometric constraints on the gradient space contour of possible solutions to the image-forming equation $I(x,y) = R(p,q)$. The algorithm discussed in this chapter provides a simple mechanism for propagating these geometric constraints.

3. EXPLOITING ADDITIONAL CONSTRAINT

Known properties of surface curvature can be used to simplify image analysis. Earlier, it was demonstrated how convexity and concavity constrain the possible matches between image points and their corresponding gradient points. Here, this analysis is extended to explore additional situations in which known properties of surface curvature can be exploited in image analysis.

Horn has considered situations in which special properties of surface photometry simplify image analysis [Horn 75], [Horn 77]. This chapter presents a complementary study. Simplifications due to special properties of surface photometry constrain the reflectance map. When viewed from great distances, the material of the maria of the moon, for example, has a reflectance function which is constant for constant cos(i)/cos(e). This results in a reflectance map whose contours of constant R(p,q) are a family of parallel straight lines. Image analysis is simplified since the base characteristics used to solve the nonlinear partial-differential equation become a predetermined family of straight lines, independent of surface topography. Here, an arbitrary reflectance map is allowed. Instead, simplifications arise from known properties of surface curvature. It is shown that image analysis simplifies for surfaces with constant image Hessian, for surfaces that are singly curved and for surfaces corresponding to objects that are generalized cones.

The results developed for these special cases deepen our understanding of the problem of computing object relief from image intensity. Relating the results developed here to the generalized cone representation of [Agin & Binford 73], [Marr 77a] and [Marr 77b] helps to delineate shape information that can be determined from object boundaries and shape

information that can be determined from shading. The results developed are of immediate practical importance since many surfaces occurring in industrial parts design and manufacture are also constrained by the surface curvature properties discussed here.

Finally, a novel technique called photometric stereo is presented. Photometric stereo exploits the additional constraint provided from a second image taken with the same imaging geometry but with a different direction of incident illumination.

3.1 SURFACES WITH CONSTANT IMAGE HESSIAN

This section examines the case for which the image Hessian matrix **H** is constant. This is useful for two reasons. First, if indeed the Hessian matrix **H** is constant over a region of the image, then local information gathered about the Hessian becomes global information. This allows one to combine local evidence about the Hessian and determine it completely. Second, if one assumes that **H** is constant over a small region of the image, then it is possible to use the intensity values present in that region to approximate surface curvature over that region.

In developing the basic image-forming equation

$$I(x,y) = R(p,q)$$

an effort has been made to stress the fact that the reflectance map is not an image. Rather, it is a convenient representation of how surface orientation determines image intensity for a particular object material and object surface, light source and viewer geometry. In this section, it will be convenient to think of the reflectance map as an image.

For the reflectance map to be an image of some object, the surface must satisfy the two differential equations:

$$\frac{\partial f(x,y)}{\partial x} = x$$

$$\frac{\partial f(x,y)}{\partial y} = y$$

The family of paraboloids:

$$z = \frac{x^2 + y^2}{2} + k \qquad (3.1.1)$$

(where k is an arbitrary constant) satisfy the two differential equations. For these surfaces, the basic image-forming equation becomes:

$$I(x,y) = R(x,y)$$

Thus, the reflectance map may be thought of as the image of a paraboloid. As a first observation, this suggests that a paraboloid would make an ideal calibration object for empirically determining a reflectance map. Typically, a sphere is used. A sphere serves almost as well if one is willing to live with a loss of accuracy as the viewing angle increases.

For the problem at hand, it is not so much of interest that the reflectance map is the image of a paraboloid, but that, for such a surface, the image Hessian matrix **H** is constant. Indeed, for the family of paraboloids described by (3.1.1), The Hessian matrix **H** is the 2 x 2 identity matrix. In general, in order to have a constant Hessian **H**, a surface with explicit representation $z = f(x,y)$ must have no terms of higher than second order (in x and y). That is, the surface can be described by the form:

$$z = ax^2 + by^2 + 2cxy + dx + ey + f \qquad (3.1.2)$$

The corresponding Hessian becomes:

$$H = 2 \begin{bmatrix} a & c \\ c & b \end{bmatrix}$$

To rule out degenerate situations, assume that the surface described by (3.1.2) is fully second order. That is, assume $ab - c^2 \neq 0$. (This assumption implies that H is nonsingular.) Subject to this assumption, (3.1.2) describes either an elliptic paraboloid or a hyperbolic paraboloid.

Suppose that image point (x_0,y_0) is known to correspond to gradient point (p_0,q_0). In order to determine the image Hessian matrix H at (x_0,y_0), it is sufficient to determine the movement in gradient space corresponding to two linearly independent movements in the image at (x_0,y_0). Horn's method for obtaining shape from shading information is based on the observation that, if $[dx,dy] = [R_p,R_q]ds$, then $[dp,dq] = [I_x,I_y]ds$ [Horn 75] [Horn 77]. (See Appendix A.1 for details). This gives one piece of information about the Hessian matrix H at (x_0,y_0). Iterating this procedure provides additional pieces of information as a characteristic strip is expanded across the image. In general, however, this information cannot be combined since the Hessian itself changes along the characteristic strip. But, if the surface is of the form (3.1.2), then the information can be combined.

For most regions of a reflectance map, the direction defined by $[R_p,R_q]$ does not change very rapidly along a characteristic strip. Nevertheless, it does not require much change to determine H to high accuracy. But note a paradox. The situation in which Horn's method achieves its greatest simplification (i.e., lunar topography) is the situation for which the base characteristics are predetermined straight lines in the image. If, in fact, the base characteristics are straight lines then it is never possible to obtain linearly independent movements in

the image by characteristic strip expansion! In such a case, it is not possible to say anything about how H behaves in directions other than the direction defined by the base characteristic, even in the cases for which H is constant over the whole image. Roughly speaking, the ability to approximate the Hessian H locally at an image point (x_0,y_0) is related to the curvature of the base characteristic passing through (x_0,y_0). If the base characteristic has high curvature at (x_0,y_0) then good conditioning, in the sense of linear independence, is achieved between closely spaced image points (i.e., over areas in the image where the approximation of constant H is likely to be a reasonable one).

There are two factors which act together to determine how the image Hessian matrix H maps a movement [dx,dy] in the image into the corresponding movement [dp,dq] in gradient space. First, in moving from an image point (x_0,y_0) to an adjacent point (x_1,y_1) there is a change in local surface orientation due to actual surface curvature. Second, because the object is viewed obliquely, the actual change in local surface orientation it altered by the foreshortening of the surface in image projection. Fortunately, it is possible to decouple surface curvature from the surface foreshortening induced by a nonzero view angle e.

Suppose image point (x,y) is known to correspond to the gradient point (p,q). Let H be the Hessian matrix H at (x,y). Define a matrix C at (x,y) by:

$$C = \cos^3(e)\begin{bmatrix} q^2+1 & -pq \\ -pq & p^2+1 \end{bmatrix} H \qquad (3.1.3)$$

Equivalently, one can solve (3.1.3) for H and obtain:

$$H = 1/\cos(e) \begin{bmatrix} p^2+1 & pq \\ pq & q^2+1 \end{bmatrix} C \qquad (3.1.4)$$

The matrix C, defined by equation (3.1.3), corresponds to an object-centered definition of surface curvature. Let $1/r_1$ and $1/r_2$ be the two eigenvalues of C with corresponding eigenvectors ω_1 and ω_2. Then r_1 and r_2 are the two principal radii of curvature of the object surface at (x,y) with corresponding directions, in the image, given by ω_1 and ω_2. Note that the values r_1 and r_2 are independent of the imaging geometry and depend only on the actual surface topography at (x,y). The first two terms of the right-hand side of equation (3.1.4) account for the "distortion" in H due to the oblique viewing angle e. The matrix C is an object-centered definition of surface curvature while the matrix H a viewer-centered definition of surface curvature. If the gradient (p,q) at image point (x,y) is known, then equations (3.1.3) and (3.1.4) make these two definitions of surface curvature equivalent.

Constraints on surface curvature are constraints on the matrix C. Equation (3.1.4) must be used to translate any such constraints into equivalent constraints on H. Some idea of the effect of surface foreshortening can be seen by examining the structure of H in more detail. First, the determinant of H is given as:

$$\det|H| = \frac{1}{\cos^4(e)} \frac{1}{r_1 r_2}$$

Roughly speaking, H contains a "scale factor" of $1/\cos^2(e)$ due to the oblique viewing angle e. More precisely, the product

$$1/\cos(e) \begin{bmatrix} p^2+1 & pq \\ pq & q^2+1 \end{bmatrix}$$

has eigenvalue $1/\cos^3(e)$ in the direction of steepest decent and eigenvalue $1/\cos(e)$ in the direction of the contour of constant e.

Elliptic and hyperbolic paraboloids are the unique surfaces for which the "increase" in H due to an increasing view angle e is precisely offset by the decrease in $1/r_1$ and $1/r_2$ due to the changing surface curvature.

3.1.1 APPROXIMATING THE IMAGE HESSIAN LOCALLY

Let us see how the assumption that H is constant over a small region of the image can be used to approximate the image Hessian at an image point (x,y) known to correspond to the gradient point (p,q). The basic observation used here is that multiplication by a constant H defines a one-to-one continuous mapping between a circle in the image centered at (x,y) and an ellipse in gradient space centered at (p,q) (see Appendix A.3). If two linearly independent directions $[dx_1,dy_1]$ and $[dx_2,dy_2]$ and the corresponding $[dp_1,dq_1]$ and $[dp_2,dq_2]$ are known, then H is determined by:

$$H = \begin{bmatrix} dp_1 & dp_2 \\ dq_1 & dq_2 \end{bmatrix} \begin{bmatrix} dx_1 & dx_2 \\ dy_1 & dy_2 \end{bmatrix}^{-1} \tag{3.1.5}$$

One correspondence can be tied down immediately. If $[dx,dy] = [R_p,R_q]ds$ then $[dp,dq] = [I_x,I_y]ds$. Thus, (3.1.5) can be rewritten as:

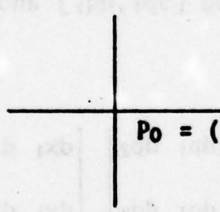
$$H = \begin{bmatrix} I_x & dp \\ I_y & dq \end{bmatrix} \begin{bmatrix} R_p & dx \\ R_q & dy \end{bmatrix}^{-1} \tag{3.1.6}$$

Equation (3.1.6) still admits an infinite number of solutions. However, for any particular second choice [dx,dy], linearly independent from $[R_p,R_q]$, the geometric constraints developed in Chapter 2 can be used

to constrain the set of [dp,dq] that can correspond to the given [dx,dy]. Each [dp,dq] not eliminated by geometric constraint defines a possible solution to (3.1.6). For each possible solution, one can proceed to sample the intensity values along the image circle centered at (x,y) and of radius $ds = \sqrt{dx^2 + dy^2}$. For each point on this circle, the corresponding reflectance value R(p+dp,q+dq), where $[dp,dq]^T = H\ [dx,dy]^T$, can also be sampled. The "best fit" Hessian at (x,y) is defined to be that choice of [dp,dq] which minimizes the least square error between the sampled intensity values and the corresponding sampled reflectance values.

Consider an example. Suppose image point $P_0 = (x_0, y_0)$ is known to correspond to gradient point (p_0, q_0) on a section of surface assumed to be convex. Then, for a particular choice of ds, the image is sampled at the four points:

$P_2 = (x_0, y_0+ds)$

$P_3 = (x_0-ds, y_0)$  $P_1 = (x_0+ds, y_0)$

$P_0 = (x_0, y_0)$

$P_4 = (x_0, y_0-ds)$

From the four intensity values thus obtained, the four reflectance map contours, corresponding to the four image points P_1, P_2, P_3 and P_4, are determined. (Note: if bounds on the eigenvalues of H are already known, then the range in gradient space used in the initial determination of the four reflectance map contours can be tightly constrained.) Figure 3-1 illustrates the four contours obtained for a particular choice of (x_0, y_0), (p_0, q_0) and ds, using the lambertian sphere example of Chapter 2.8. No a priori constraint is assumed on the magnitude of the eigenvalues of H.

Convexity allows us to assert the two inequalities:

$$R_p I_x + R_q I_y \geq 0$$

$$dpdx + dqdy \geq 0$$

After applying these two inequalities to each of the four contours of figure 3-1, the constrained contours of figure 3-2 are obtained.

There is one additional constraint available. **H** defines a linear transformation. For the particular choice of P_1, P_2, P_3 and P_4 above, observe that the [dx,dy] used to move from (x_0,y_0) to P_1 is the negative of the [dx,dy] used to move from (x_0,y_0) to P_3. (Similarly, for P_2 and P_4.) Thus the corresponding [dp,dq]'s must be the negative of each other. In particular, the magnitude of the [dp,dq]'s must be equal and they must be π radians out of phase. Each point in the contour of possible (p,q) for P_1 which does not have a match in P_3 whose distance from (p_0,q_0) is the same and whose angular position with respect to (p_0,q_0) is π radians out of phase can be excluded. Similarly, each point in the contour of possible (p,q) for P_3 which does not have a match in P_1 whose distance from (p_0,q_0) is the same and whose angular position with respect to (p_0,q_0) is π radians out of phase can be excluded. (Note: these criteria must be applied with caution otherwise all points will be excluded due to the fact that **H** is not really constant.) The identical constraint holds between P_2 and P_4. Figure 3-3 shows the contour remaining when the above criteria are applied to exclude "obvious" bad matches.

Of the two pairs of contour, corresponding to P_1 and P_3 and to P_2 and P_4, select the pair that is closest to being $\pi/2$ radians out of phase with the direction $[I_x,I_y]$. (This choice achieves superior conditioning in the estimation of **H**.) For each [dp,dq] remaining in the most highly constrained contour from the selected pair, use (3.1.6) to estimate **H**.

Figure 3-4 shows the family of ellipses generated by each point remaining on the contour for P_4. The vector superimposed on figure 3-4 points in the direction $[I_x, I_y]$. Figure 3-5 shows least squares "best fit" Hessian matrix at (x_0, y_0).

The "best fit" Hessian matrix of figure 3-5 has done a good job of solving for the curvature of the object surface at (x_0, y_0). It is important, however, to be precise about what it is that has been computed. There are two assumptions underlying the computation. First, it is assumed that the Hessian matrix **H** is constant over the area in the image determined by the circle centered at (x_0, y_0) and of radius ds. Second, it is assumed that the gradient point (p_0, q_0) corresponding to image point (x_0, y_0) is known. Equation (3.1.3) tells how to choose ds to compensate for the foreshortening due to an oblique viewing angle e and thus achieve uniform sampling resolution over the entire surface. Actually, by sampling equally spaced points around the image circle, different weights are applied to different directions along the surface. A more sophisticated sampling scheme is required to remove this difficulty. A more serious restriction, however, is related to the second assumption. If the center of the ellipse has not been accurately positioned then a least squares "best fit" is akin to groping around in the dark. This example illustrates not so much a useful method for finding the gradient points corresponding to a set of image points but the fact that the intensity values in an image can be used locally to constrain the behavior of the image Hessian matrix **H** in directions other than those defined by the current $[I_x, I_y]$.

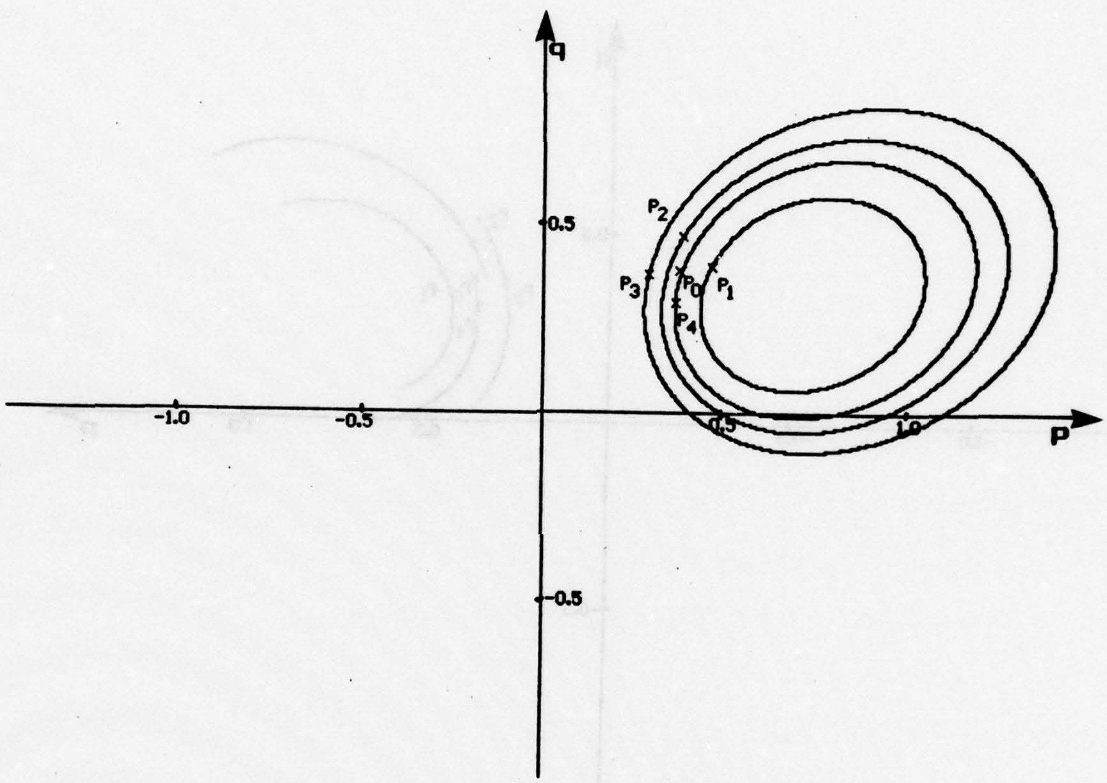


Figure 3-1 The four reflectance map contours of possible gradients corresponding to the four sampled image points P_1, P_2, P_3 and P_4. Crosses mark the exact location, determined analytically, of P_0, P_1, P_2, P_3 and P_4.

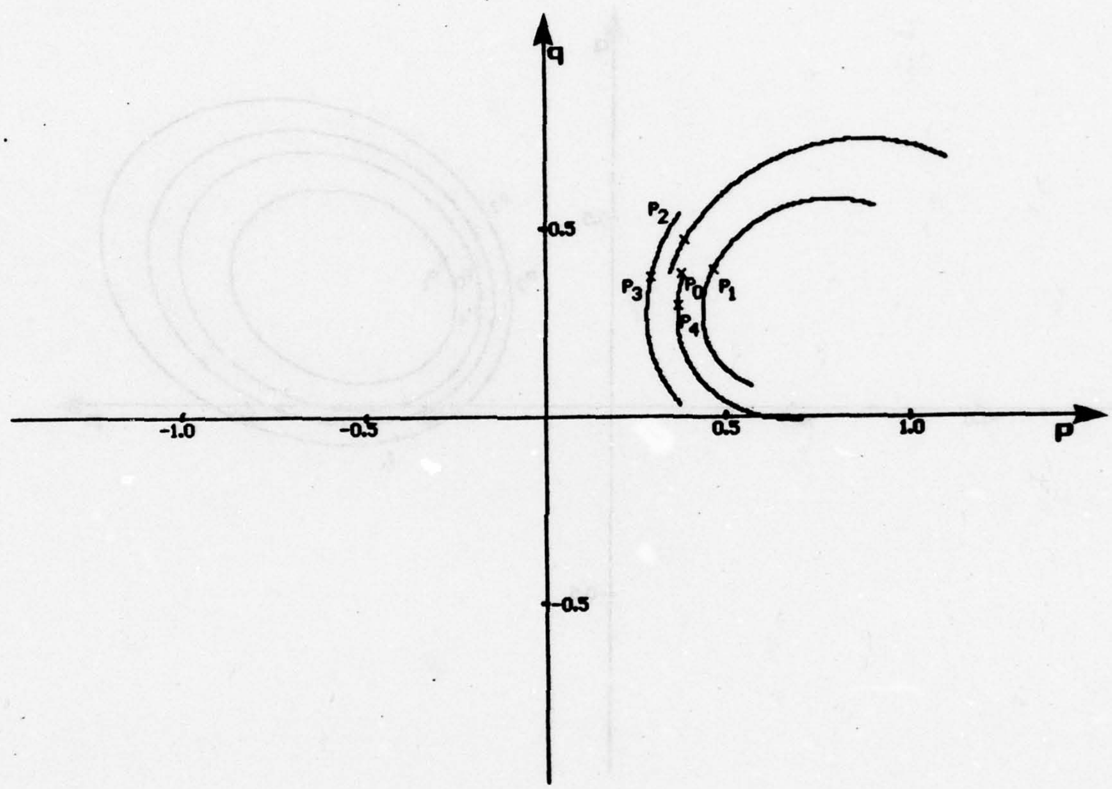


Figure 3-2 **The restricted subsection of contour for P_1, P_2, P_3 and P_4 obtained if the section of surface under consideration is convex.**

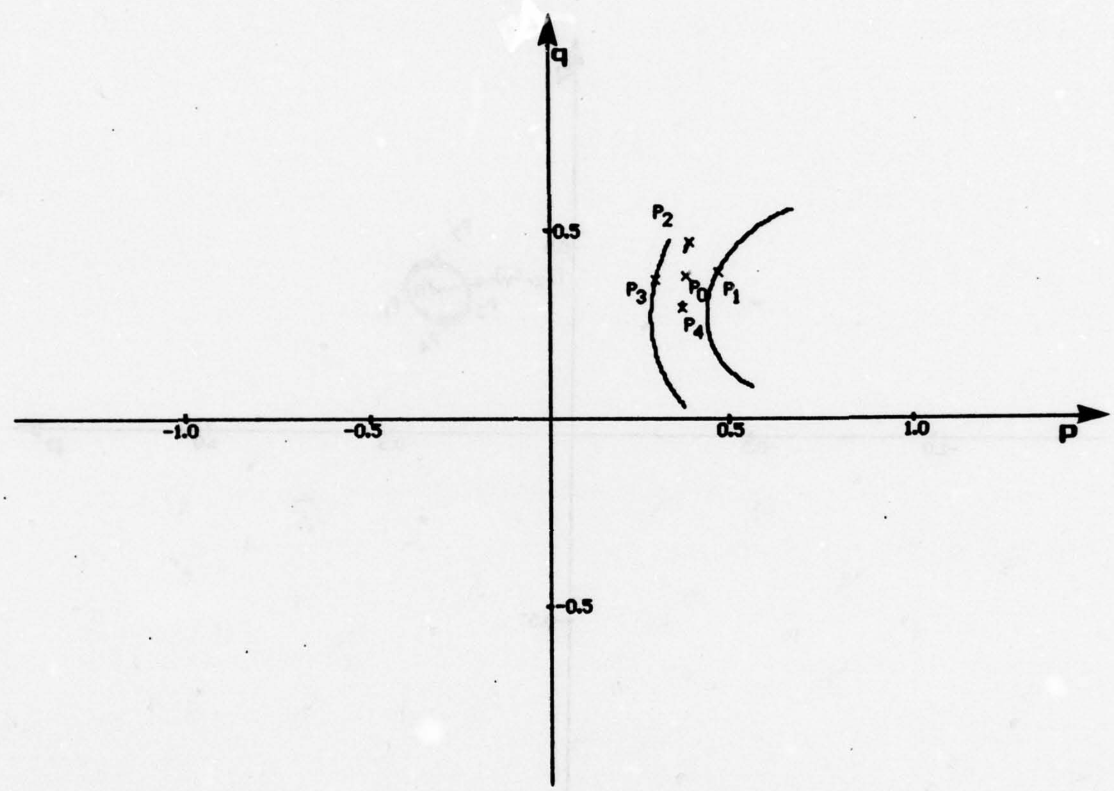


Figure 3-3 The restricted subsection of contour for P_1, P_2, P_3 and P_4 obtained when "linearity" is used to exclude obvious bad matches.

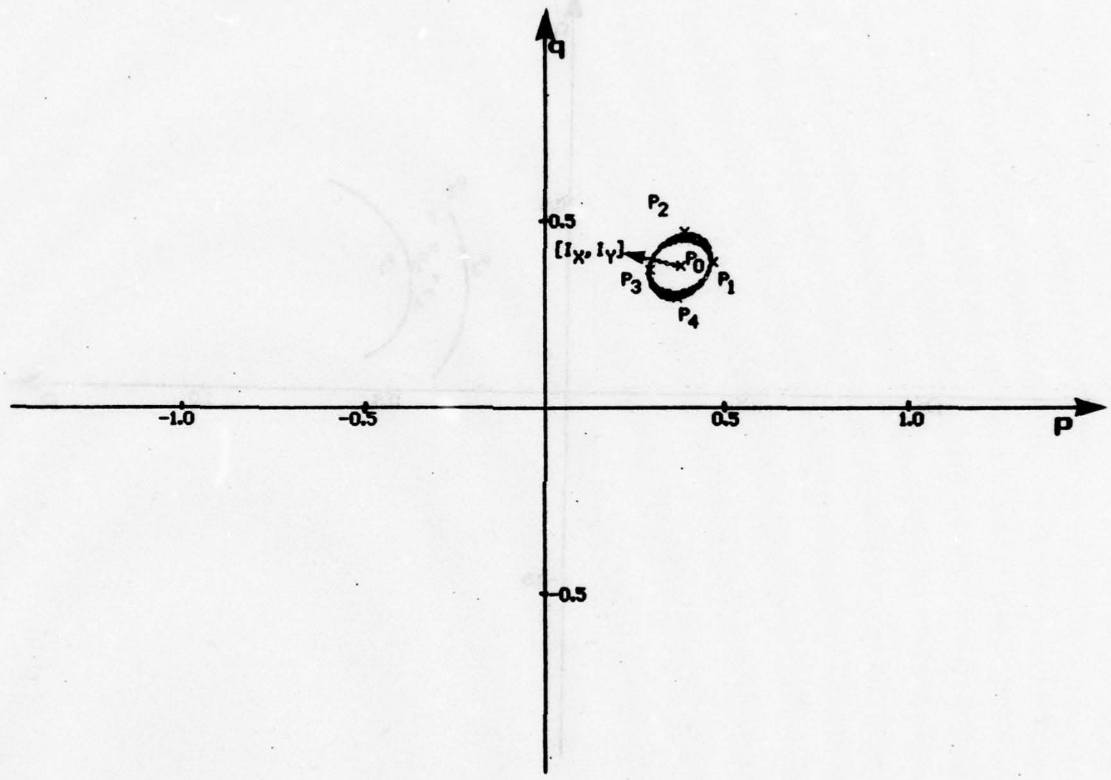


Figure 3-4 The family of ellipses which characterize the uncertainty remaining in the determination the image Hessian H. The vector at P_0 is in the direction of the normal to the contour of constant intensity in the image at (x_0, y_0)

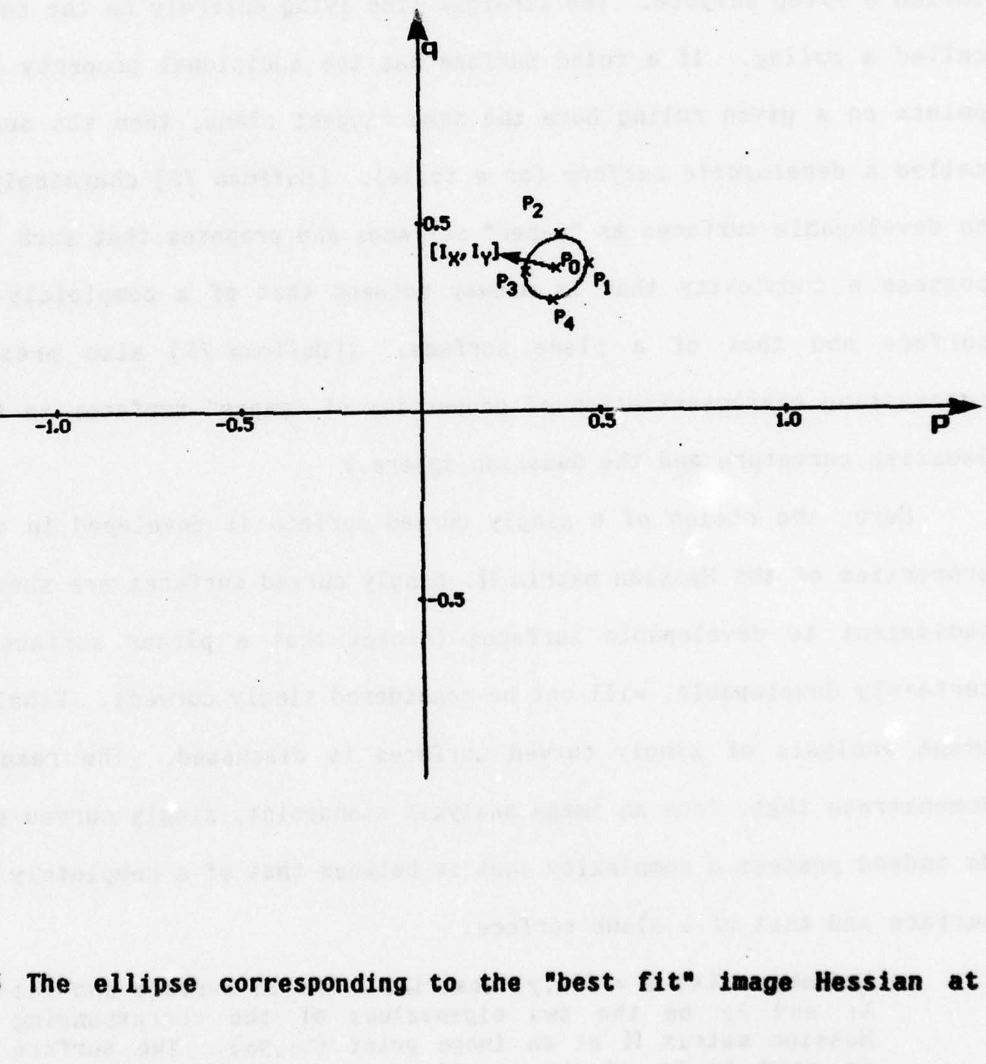


Figure 3-5 The ellipse corresponding to the "best fit" image Hessian at (x_0, y_0).

3.2 SINGLY CURVED SURFACES

An important class of surfaces in differential geometry are those which have the property that, through every point on the surface, there passes at least one straight line lying entirely on it. Such a surface is called a *ruled surface*. The straight line lying entirely on the surface is called a *ruling*. If a ruled surface has the additional property that all points on a given ruling have the same tangent plane, then the surface is called a *developable surface* (or a *torse*). [Huffman 75] charmingly refers to developable surfaces as "paper" surfaces and proposes that such surfaces possess a complexity that is midway between that of a completely general surface and that of a plane surface. ([Huffman 75] also presents an interesting characterization of properties of "paper" surfaces in terms of Gaussian curvature and the Gaussian sphere.)

Here, the notion of a singly curved surface is developed in terms of properties of the Hessian matrix H. Singly curved surfaces are shown to be equivalent to developable surfaces (except that a planar surface, while certainly developable, will not be considered singly curved). Finally, the image analysis of singly curved surfaces is discussed. The result will demonstrate that, from an image analysis standpoint, singly curved surfaces do indeed possess a complexity that is between that of a completely general surface and that of a plane surface.

Definition. Let $z = f(x,y)$ describe a smooth surface and let λ_1 and λ_2 be the two eigenvalues of the corresponding Hessian matrix H at an image point (x_0,y_0). The surface is said to be *singly curved at* (x_0,y_0) if and only if exactly one of λ_1 and λ_2 is equal to zero. The surface is said to be *singly curved* if it is singly curved at each image point (x,y) on the surface.

This viewer-centered definition of a singly curved surface can be related to the object-centered principal radii of curvature. From equations (3.1.3) and (3.1.4) above, it can be shown that one of the eigenvalues λ_1 and λ_2 of H is zero if and only if one of the principal radii of curvature r_1 and r_2 is infinite (see Appendix A.5).

For surfaces expressed in the form $z = f(x,y)$, the equation:

$$\frac{\partial^2 f(x,y)}{\partial^2 x}\,\frac{\partial^2 f(x,y)}{\partial^2 y} = \left[\frac{\partial^2 f(x,y)}{\partial x \partial y}\right]^2 \qquad (3.2.1)$$

is the differential equation characterizing developable surfaces. In the current notation, this equation is equivalent to the equation:

$$\det|\mathbf{H}| = 0 \qquad (3.2.2)$$

Equation (3.2.2) is satisfied at an image point (x,y) if and only if at least one of the eigenvalues of the corresponding H is zero at (x,y). Thus, (3.2.1) is satisfied for all surface points (x,y) if and only if at least one of the eigenvalues of H is zero at each image point (x,y). If both eigenvalues of H are zero for all image points (x,y), then the corresponding equation $z = f(x,y)$ describes a plane. For convenience, assume that at least one of λ_1 and λ_2 is nonzero at each image point (x,y). Thus, a nonplanar surface $z = f(x,y)$ is developable if and only if it is singly curved.

Let $z = f(x,y)$ be a singly curved surface. Suppose that a point (x_0,y_0) in the image is known to correspond to a point (p_0,q_0) in gradient space. Then, the Hessian matrix H at (x_0,y_0) is completely determined. Indeed, H is given as the matrix product:

$$\mathbf{H} = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix} \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{bmatrix} \qquad (3.2.3)$$

where
$$\lambda = \frac{\sqrt{I_x^2 + I_y^2}}{R_p \cos(\alpha) + R_q \sin(\alpha)}$$

and
$$\tan(\alpha) = \frac{I_y}{I_x}$$

As before, I_x and I_y denote the first partial derivatives of $I(x,y)$ at image point (x_0,y_0) and R_p and R_q denote the first partial derivatives of $R(p,q)$ at the corresponding gradient point (p_0,q_0).

Thus, given any initial image point (x_0,y_0) known to correspond to gradient point (p_0,q_0), the Hessian matrix of $z = f(x,y)$ at (x_0,y_0) is determined by (3.2.3). The Hessian matrix H so determined can then be used to find the new gradient corresponding to an arbitrary movement $[dx,dy]$ in the image according to the equation:

$$[dp,dq]^T = H\,[dx,dy]^T \tag{3.2.4}$$

The operations embodied in (3.2.3) and (3.2.4) above can be iterated to trace out an arbitrary family of curves on the surface. For singly curved surfaces, one is not confined to tracing out the characteristics of Horn's original method for obtaining shape from shading information [Horn 75], [Horn 77].

This result should not be terribly surprising. The fact that H has one zero eigenvalue means that there is one direction of movement in the image which results in no change to surface orientation. The orthogonal direction α is determined by the vector $[I_x,I_y]$. The component of any movement $[dx,dy]$ perpendicular to $[I_x,I_y]$ is in the direction of a ruling on the developable surface $z = f(x,y)$ and thus does not cause a change to the gradient (p,q). The component of $[dx,dy]$ in the direction $[I_x,I_y]$ causes a change $[dp,dq]$ to the gradient in the direction α where the "scale factor" for that change is given by the value of λ.

The points in gradient space corresponding to points on an arbitrary singly curved surface $z = f(x,y)$ are constrained to lie on a one-parameter curve in gradient space. This is just another manifestation of the observation that singly curved surfaces possess a complexity midway between that of a plane surface, where surface points map into a single point in gradient space, and that of a completely general surface, where surface points map into a two-parameter region in gradient space.

Figure 3-6 is the image of a right circular cone of base radius b and height h generated using the reflectance map of figure 2-4. (For this example, $h = 2b$.) It can be shown that the points in gradient space corresponding to points on a right circular cone lie on the one-parameter curve in gradient space given parametrically by:

$$p = \tan(t) \tag{3.2.5}$$

$$q = \frac{b}{h}\frac{1}{\cos(t)} \tag{3.2.6}$$

where $-\pi/2 < t < \pi/2$. The parameter t has a physical interpretation. The circular cross-section of the cone can be represented, in cylindrical coordinates, by the function $\rho(\theta) = 1$, where θ measures angular position about the y-axis. If θ is chosen so that $\theta = 0$ points in the direction of the viewer, then the parameter t in (3.2.5) and (3.2.6) is this angle θ.

Figure 3-7 shows the curve in gradient space determined by the parametric equations (3.2.5) and (3.2.6) superimposed on the reflectance map used to generate the image of figure 3-6. There exists a one-to-one continuous mapping between any horizontal image intensity profile from figure 3-6 and the curve in gradient space determined by (3.2.5) and (3.2.6). Thus, finding the point (p_0,q_0) in gradient space corresponding to any image intensity point $I(x_0,y_0) = \alpha$ from figure 3-6 simplifies to the

problem of determining reflectance map values on the curve given by (3.2.5) and (3.2.6) for which $R(p,q) = \alpha$. If an intensity profile is scanned sequentially from left to right then possible multiple solutions can be resolved by chosing the solution which is "closest" to the previous solution in the direction of increasing t. This follows as a consequence of the general fact that all 1-1 continuous mappings are monotonic and the particular observation, for this case, that the mapping is monotonically nondecreasing.

3.3 GENERALIZED CONES

The previous section considered the imaging of singly curved surfaces. This section extends that work to consider a more general class of surfaces called generalized cones. Generalized cones are doubly curved. The curvature of a generalized cone, however, conveniently decouples, in an object-centered representation of shape. For appropriate viewing conditions, this decoupling carries over to images of generalized cones. These images can be analyzed "almost" as if the surface were singly curved. The first result of this section is to demonstrate that, if the cone axis is parallel to the viewing plane, then images of right generalized cones with circular cross-section can be analyzed exactly as if the surface were singly curved. In this case, the additional information required to account for the second degree of freedom in curvature is not embedded in the intensity values present in the image but rather in the object silhouette. The more general case, in which the cone axis makes a nonzero angle with the viewing plane, is more difficult to analyze. The same decoupling of curvature holds but this decoupling is more difficult to recover in an image. Finally, a few results are given for the case in

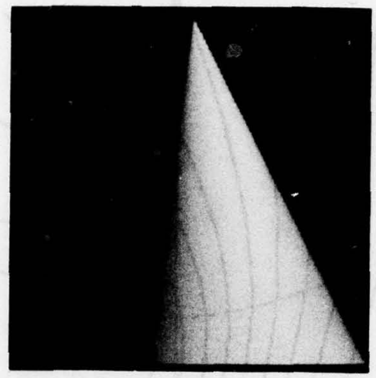


Figure 3-6 The synthesized image of a right circular cone, with height equal to twice the base radius and axis parallel to the image plane. The surface reflectance is lambertian with a single distant light source at gradient point p_s = 0.7 and q_s = 0.3.

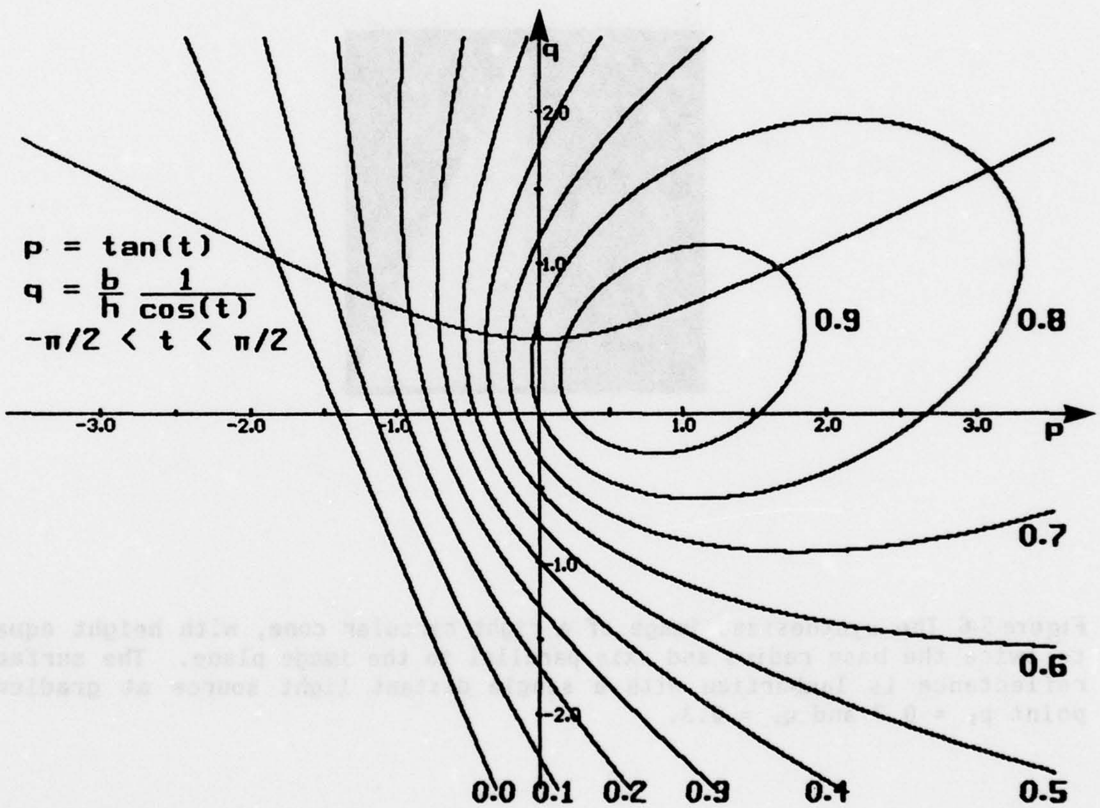


Figure 3-7 The one parameter curve in gradient space corresponding to points on the cone of figure 3-6.

which the cross-section is allowed to be an arbitrary convex function.

This section is important for two reasons. First, it extends the approach taken in the previous section to a broader class of surfaces. Second, it is an initial attempt to construct a bridge between methods of determining surface topography by analyzing the intensity variation across smooth sections of object surface and methods for determining object shape by analyzing the occluding contours present in an image.

This section is preliminary. The question asked is, "If the object is a generalized cone, then how does this constrain the intensity values recorded in an image of the object?" The question one really wants to ask is, "How can the intensity values recorded in an image be used to determine whether the object is a generalized cone and, if so, to determine its axis and cross-section function?" At best, the results here correspond to a method for checking the consistency of hypotheses about surface shape determined from an analysis of the object silhouette against the actual intensity values recorded over smooth sections of the object surface.

The concept of a generalized cone has its genesis in the generalized cylinder representation of [Agin & Binford 73]. There, generalized cylinders were used as convenient representation scheme for describing complex shapes. Generalized cones arise in the work of [Marr 77a], [Marr 77b]. Here, the generalized cone emerges not so much as a convenient representation scheme but rather as an interpretation that is forced if one tries to develop a theory of how to infer the shape of objects from their silhouettes.

A *generalized cone* is defined to be the surface swept out by moving a simple smooth cross-section $\rho(\theta)$ along a straight axis Λ, at the same time magnifying or contracting it in a smoothly varying way. Let $h(\lambda)$ be the

axial scaling function where λ denotes distance along the Λ axis. The angle ψ between the axis Λ and a plane containing a cross-section is called the eccentricity of the cone. For the present, it will be convenient to add two additional simplifying assumptions. First, assume that the eccentricity $\psi = \pi/2$. With this first assumption, the cone is called a *right generalized cone*. Second, assume that the cross-section is circular. Without loss of generality, one can further assume that the axis Λ passes through the center of the circle. With this second assumption, the cone is called a *right generalized cone with circular cross-section.*

[Marr 77a] has defined methods for finding the projection of the axis Λ on the image plane of a generalized cone from an analysis of its occluding contour. The goal here is to provide a complementary study to interpret the intensity values recorded from the interior smooth sections of a generalized cone.

Let us begin with one further simplification. Assume that the axis Λ is parallel to the image plane. Since a rotation of object space induces an equal rotation in gradient space, one can, without loss of generality, assume that the axis Λ coincides with the image y-axis. This last assumption about the viewing direction will allow for the convenient decoupling of the generalized cone's curvature so that the image analysis problem becomes equivalent to that of analyzing a singly curved object. Distance along the axis Λ is equal to distance along the image y-axis so that the axial scaling function can be denoted as $h(y)$. Let the circular cross-section function be denoted by $\rho(\theta) = 1$, where $\theta = 0$ points in the direction of the viewer. It can be shown that the points in gradient space corresponding to points on such a right generalized cone with circular cross-section lie in the two-parameter region in gradient space given

parametrically by:

$$p = \tan(\theta) \tag{3.3.1}$$

$$q = \frac{-h'(y)}{\cos(\theta)} \tag{3.3.2}$$

where $h'(y)$ denotes the derivative of $h(y)$ with respect to y and θ varies between $-\pi/2$ and $\pi/2$.

Figure 3-6 was an example of a right generalized cone with circular cross-section. A right circular cone of base radius *b* and height *h* has axial scaling function $h(y)$ where:

$$h'(y) = \frac{-b}{h}$$

In general, (3.3.1) and (3.3.2) define a two parameter region in gradient space. One of the parameters is θ and the other is $h'(y)$. Of course, if $h'(y)$ is constant, as in figure 3-6, then the surface is singly curved. The interesting observation, however, is that when the axis Λ is parallel to the viewing plane, the value of $h'(\lambda)$ can always be determined directly from the boundary contour. For a particular value of y, the curve in gradient space generated by (3.3.1) and (3.3.2) is a scaled version of the curve illustrated in figure 3-7. Differing values of $-h'(y)$ introduce a different scale factor in q. Thus, finding the point (p_0, q_0) in gradient space corresponding to an image intensity point $I(x_0, y_0) = \alpha$ simplifies to a two step process. First, for the particular value of y_0, determine $h'(y_0)$ as the rate of change of object radius with respect to y, or equivalently as one half the rate of change of object diameter with respect to y, at the object boundary points along the image profile $y = y_0$. Second, as in the case of a singly curved object, scan a horizontal intensity

profile to determine the correct reflectance value on the curve given by (3.3.1) and (3.3.2).

Figure 3-8 is a more general example. Here, the axial scaling function is a sinusoid while the cross-section function remains circular. The surface depicted in figure 3-8 is doubly curved. Yet, the curvature decouples so that, from a image analysis point of view, the surface behaves as if it were singly curved. Figure 3-9 superimposes a collection of these curves on the reflectance map used to generate figure 3-8. (Note: the sphere example of Chapter 2.8 is also a right generalized cone with circular cross-section with an axis Λ that can always be chosen parallel to the viewing plane!)

Now consider the case in which the axis Λ of a right generalized cone with circular cross-section is not parallel to the viewing plane. Once again, a rotation of object space induces an equal rotation in gradient space so that, without loss of generality, assume that the projection of the cone's axis Λ on the image plane coincides with the image y-axis. Let the angle between the Λ axis and the viewing plane be ϕ (measured so that positive ϕ implies that the Λ axis is tilted towards the viewer). In this situation, distance along the image y-axis is simply a foreshortened version of distance along Λ. In particular,

$$y = \lambda \cos(\phi)$$

so that the derivative of the axial scaling function $h(\lambda)$ is given by:

$$h'(\lambda) = \cos(\phi)\ h'(y) = \cos(\phi)\ \frac{\text{rate of change of image diameter}}{2}$$

Now, tilting object space an angle ϕ about the x-axis takes gradient point (p,q) onto gradient point (p,',q') where:

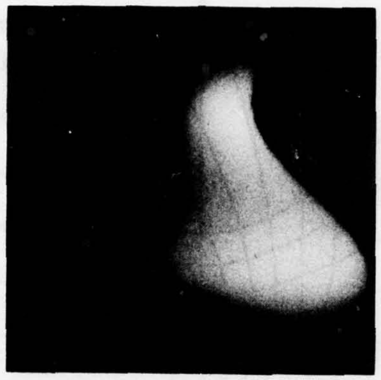


Figure 3-8 The synthesized image of a right generalized cone with circular cross-section and axis Λ parallel to the image plane. The axial scaling function h(y) is a sinusoid. The surface reflectance is lambertian with a single distant light source at gradient point $p_s = 0.7$ and $q_s = 0.3$.

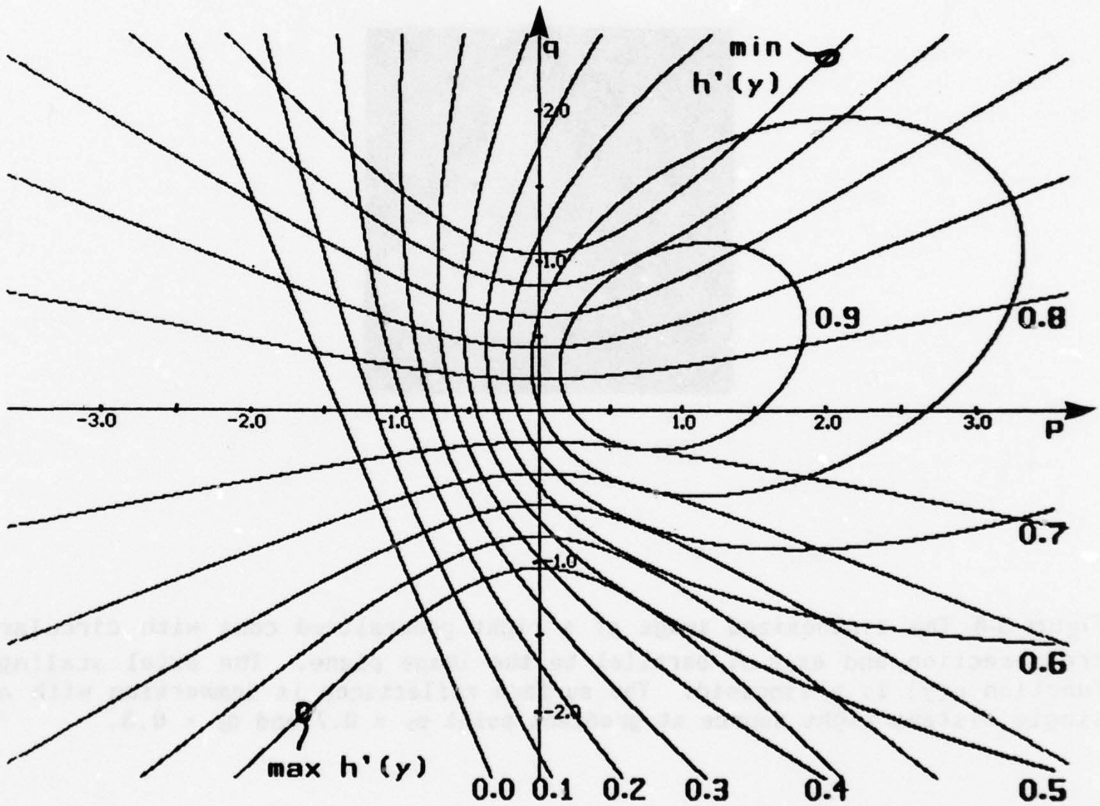


Figure 3-9 The region in gradient space corresponding to points on the right generalized cone of figure 3-8, plotted as a family of curves. Note that the region in gradient space lies below the curve determined by the minimum value of h'(y) and above the curve determined by the maximum value of h'(y).

$$p' = \frac{p}{\cos(\phi) + q\sin(\phi)}$$

$$q' = -\frac{\sin(\phi) - q\cos(\phi)}{\cos(\phi) + q\sin(\phi)}$$

The points in gradient space corresponding to points on a right generalized cone with circular cross-section lie in the two-parameter region in gradient space given parametrically by:

$$p = \frac{\sin(\theta)}{\cos(\phi)[\cos(\theta) - \sin(\phi)h'(y)]} \tag{3.3.3}$$

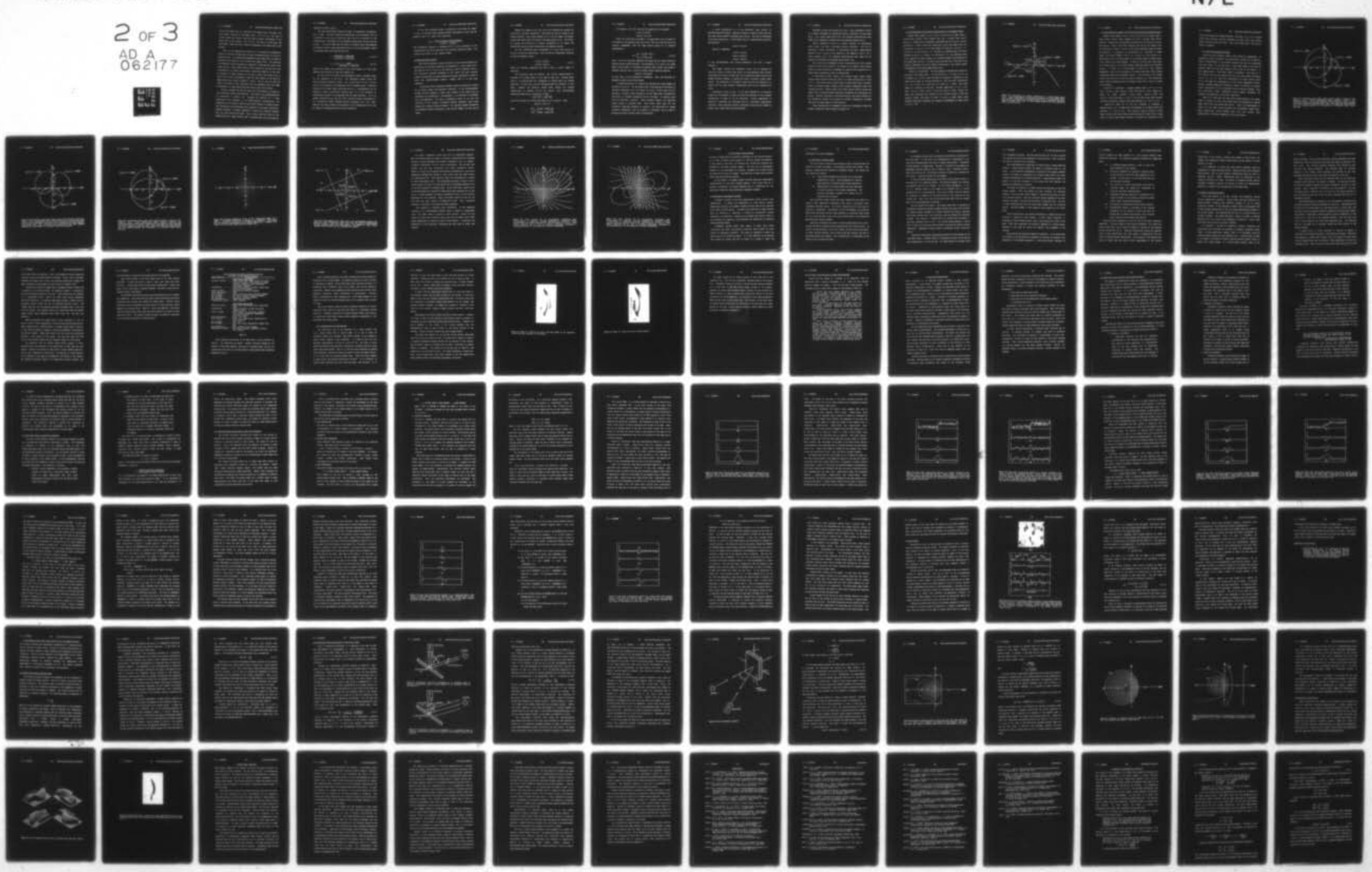
$$q = -\tan(\phi) + \frac{-h'(y)}{\cos(\phi)[\cos(\theta) - \sin(\phi)h'(y)]} \tag{3.3.4}$$

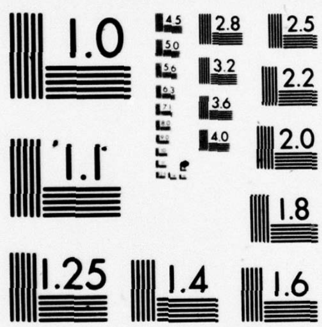
where $h'(y)$ denotes the derivative of $h(y)$ with respect to y and $-\pi/2 < \theta < \pi/2$. For a particular value of y, the points corresponding to a cross-section on a right generalized cone with circular cross-section again lie on a one-parameter curve in gradient space. If ϕ is known, then one can proceed as before, with only a slight complication in the mathematical expressions required. In general, however, the value of ϕ is unknown. This leads to two difficulties. First, without knowing ϕ, the curve in gradient space generated by (3.3.3) and (3.3.4) cannot be determined. Second, without knowing ϕ, the intensity profile in the image corresponding to this curve cannot be determined.

Consider a fixed value of $y = y_0$. Let (x_L, y_0) denote the "left" boundary point and (x_R, y_0) denote the "right" boundary point determined by the intersection of the line $y = y_0$ with the image silhouette. By the way the image axes have been aligned with respect to the cone, it must be the case that $x_L = -x_R$. Now, the image intensity profile corresponding to the curve generated by (3.3.3) and (3.3.4) lies on the ellipse centered at

AD-A062 177 MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTE-ETC F/G 20/6
REFLECTANCE MAP TECHNIQUES FOR ANALYZING SURFACE DEFECTS--ETC.(U)
JUN 78 R. J. WOODHAM N00014-77-C-0389
UNCLASSIFIED AI-TR-457 N/L

2 OF 3
AD A
062177





MICROCOPY RESOLUTION TEST CHART

$(0, y_0)$ with major axis x_R, parallel to the image x-axis, and minor axis $|x_R \sin(\phi)|$, parallel to the image y-axis. If $\phi > 0$ (i.e., the cone is tilted toward the viewer), then the correct profile corresponds to the lower half of the ellipse. If $\phi < 0$ (i.e., the cone is tilted away from the viewer), then the correct profile corresponds to the upper half of the ellipse.

Thus, for any hypothesized value of ϕ, it is possible to determine a family of profiles in image space and the corresponding family of curves in gradient space. The gradient corresponding to each image point can then be determined once again as if the surface were singly curved. The gradients so determined can be integrated to determine a range profile across each image profile. These range profiles can be analyzed to see how well they correspond to a circular cross-section, tilted by an angle ϕ. In this way, an analysis of image intensity can be used to verify the choice of ϕ. A "best fit" ϕ can be determined in much the same fashion as the best-fit image Hessian was found in Chapter 3.1 above. That is, one can find the value of ϕ which is most consistent with the hypothesis that the image corresponds to a right generalized cone with circular cross-section.

Finally, consider the case for which the cross-section is allowed to be an arbitrary smooth convex function. This result to be established here is essentially a negative one. Although the curvature of a right generalized cone decouples in its object-centered representation, the assumption of an arbitrary convex cross-section function no longer permits that decoupling to be determined from an analysis of the object silhouette. Roughly speaking, it becomes impossible to resolve the trade-off between axis tilt and cross-section shape. This, in turn, does not allow for the simplification in image analysis that is achieved when the cross-section

function is known a priori.

Let the cross-section function be given, in cylindrical coordinates, by $r = \rho(\theta)$. To avoid unecessary complication, this discussion will be restricted to the case in which the cone's axis Λ is parallel to the viewing plane. Once again, let the smoothly varying axial scaling function be $h(\lambda)$. It can be shown that the points in gradient space corresponding to points on such a right generalized cone lie in the two-parameter region in gradient space given parametrically by:

$$p = \frac{\rho(\theta)\sin(\theta) - \rho'(\theta)\cos(\theta)}{\rho(\theta)\cos(\theta) + \rho'(\theta)\sin(\theta)} \tag{3.3.5}$$

$$q = -h'(\lambda)\frac{\rho(\theta)}{\rho(\theta)\cos(\theta) + \rho'(\theta)\sin(\theta)} \tag{3.3.6}$$

where $h'(\lambda)$ denotes the derivative of $h(\lambda)$ with respect to λ and $\rho'(\theta)$ denotes the derivative of $\rho(\theta)$ with respect to θ.

(3.3.5) and (3.3.6) define the two-parameter region in gradient space in the parameters θ and $h'(\lambda)$. The decoupling between these two object-centered parameters remains. For a fixed value of $h'(\lambda)$, (3.3.5) and (3.3.6) determine a one-dimensional curve in gradient space. Differing values of $h'(\lambda)$ again simply scale this curve in q. Unfortunately, for an arbitrary convex cross-section function $\rho(\theta)$, it is not possible to conveniently discover this decoupling from the image silhouette. It is no longer necessarily true that the "left" occluding contour corresponds to $\theta = -\pi/2$ and the "right" occluding contour corresponds to $\theta = \pi/2$. Thus, even for the case in which Λ lies parallel to the viewing plane, $h'(\lambda)$ can no longer be determined directly from the silhouette.

If the left occluding contour corresponds to the value θ_L (where $-\pi < \theta_L < 0$) and the right occluding contour corresponds to the value θ_R (where $0 < \theta_R < \pi$). Then, $h'(\lambda)$ is given by:

$$h'(\lambda) = \frac{\text{rate of change of object diameter}}{[\sin(-\theta_L) + \sin(\theta_R)]}$$

The information required to determine $h'(\lambda)$ is still contained in the silhouette but it can be used only if the values θ_L and θ_R are known. The values of θ_L and θ_R, in turn, depend on $\rho(\theta)$.

3.4 PHOTOMETRIC STEREO

This section develops a useful alternative to the previous methods for determining the surface orientation corresponding to a *given image point*. The equation $I(x,y) = R(p,q)$ is *one* equation in the two unknowns p and q. The theoretical machinery developed in this report has been oriented towards seeking ways to exploit additional curvature constraint in order to solve this underdetermined equation. Another idea is to get more equations.

Traditional stereo techniques determine range by relating two images of an object viewed from different directions. If the correspondence between picture elements is known, then distance to the object can be calculated by triangulation. Unfortunately, it is difficult to determine this correspondence. The idea of photometric stereo is to vary the direction of the incident illumination between successive views while holding the viewing direction constant. This provides enough information to determine surface orientation at each picture element, as will now be shown.

Suppose two images $I_1(x,y)$ and $I_2(x,y)$ are obtained by varying the direction of incident illumination. Since there has been no change in the imaging geometry, each picture element (x,y) in the two images corresponds to the same object point and hence to the same gradient (p,q). The effect of varying the direction of incident illumination is to change the reflectance map $R(p,q)$ that characterizes the imaging situation.

Let the reflectance maps corresponding to $I_1(x,y)$ and $I_2(x,y)$ be $R_1(p,q)$ and $R_2(p,q)$ respectively. There are now two independent equations in the two unknowns p and q:

$$\begin{aligned} I_1(x,y) &= R_1(p,q) \\ I_2(x,y) &= R_2(p,q) \end{aligned} \tag{3.4.1}$$

These two nonlinear equations will have at most a finite number of solutions.

Two reflectance maps are required. But, further simplification is obtained if the phase angle g is the same in each view. Constant phase angle g is achieved when the illumination is rotated about the viewing direction. In this case, the two reflectance maps are rotations of each other. Suppose two identical distant point sources are located respectively at gradient (p_1,q_1) and (p_2,q_2), with corresponding reflectance maps $R_1(p,q)$ and $R_2(p,q)$, so that

$$\sqrt{p_1^2 + q_1^2} = \sqrt{p_2^2 + q_2^2}$$

Let θ be the angle of rotation that takes (p_1,q_1) to (p_2,q_2). Then,

$$R_2(p,q) = R_1(p',q')$$

where

$$\begin{bmatrix} p' \\ q' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix}$$

For example, if $\theta = 90^\circ$, then the two equations (3.4.1) become:

$$I_1(x,y) = R_1(p,q)$$

$$I_2(x,y) = R_1(-q,p)$$

There are three ways to proceeed. If image point (x,y) is known to correspond to the gradient point (p,q) and if $[R_{1p}, R_{1q}]$ and $[R_{2p}, R_{2q}]$ are linearly independent, then the image Hessian matrix H is uniquely determined as:

$$H = \begin{bmatrix} I_{1x} & I_{2x} \\ I_{1y} & I_{2y} \end{bmatrix} \begin{bmatrix} R_{1p} & R_{2p} \\ R_{1q} & R_{2q} \end{bmatrix}^{-1} \qquad (3.4.2)$$

(Here, the first subscript identifies the image and the second subscript denotes partial differentiation.) One possibility is to start at an image point (x_0, y_0) known to correspond to gradient point (p_0, q_0) and expand a complete solution over a section of smooth surface using the equation

$$[dp, dq]^T = H\,[dx, dy]^T$$

At each step, (3.4.2) determines H completely. Thus, the direction of movement [dx,dy] can be freely chosen.

A second possibility is to explicitly determine the finite number of solutions to (3.4.1). For the special case of the material of the maria of the moon, equations (3.4.1) are linear in p and q. In this case, (3.4.1) determine a unique surface orientation at each image point, provided the directions of incident illumination are not collinear. In general, more than one solution is possible. These solutions, however, are typically widely spaced in gradient space. They provide ideal input for the relaxation algorithm presented in Chapter 2. The application of simple curvature assumptions, such as convexity or concavity, can be used to quickly nail down a unique surface interpretation.

A third possiblility is to use additional light sources to overdetermine the solution. Suppose the object is viewed a third time with the light source placed at gradient point (p_3,q_3). Call the corresponding reflectance map $R_3(p,q)$. A third image $I_3(x,y)$ is obtained which satisfies the equation:

$$I_3(x,y) = R_3(p,q)$$

The set of equations:

$$I_1(x,y) = R_1(p,q)$$

$$I_2(x,y) = R_2(p,q)$$

$$I_3(x,y) = R_3(p,q)$$

is now overdetermined, and, barring degeneracy, will have a unique solution.

The images required for photometric stereo can be obtained by explicitly moving a single light source, by using multiple light sources calibrated with respect to each other or by rotating the object surface and imaging hardware together to simulate the effect of moving a single light source. The equivalent of photometric stereo can also be achieved in a single view by using multiple illuminations which can be separated by color.

Photometric stereo is fast. It has been developed as a practical scheme for environments in which the nature and position of the incident illumination is known or can be controlled. Initial computation is required to determine the reflectance map for a particular experimental situation. Once calibrated, however, photometric stereo can be reduced to simple table lookup and/or search operations.

Photometric stereo can be used in two ways. First, photometric stereo is a general technique for determining surface orientation at each image point. For a given image point (x,y), the equations characterizing each image can be combined to determine the corresponding gradient (p,q).

Second, photometric stereo is a general technique for determining object points that have a particular surface orientation. This use of photometric stereo corresponds to interpreting the basic image-forming equation $I(x,y) = R(p,q)$ as one equation in the unknowns x and y. For a given gradient (p,q), the equations characterizing each image can be combined to determine corresponding object points (x,y).

This latter use of photometric stereo is appropriate for the so called industrial bin-of-parts problem. The location in an image of "key" object points is often sufficient to determine the position and orientation of a known object tossed onto a table or conveyor belt.

A particularly useful special case concerns object points whose surface normal directly faces the viewer (i.e. object points with $p = 0$ and $q = 0$). Such points form a unique class of image points whose intensity value is invariant under rotation of the incident illumination about the viewing direction. Object points with surface normal directly facing the viewer can be located without explicitly determining the reflectance map R(p,q). Whatever the nature of the function R(p,q), the value of R(0,0) is not changed by varying the direction of incident illumination, provided only that the phase angle g is held constant.

These applications of photometric stereo are illustrated using the simple, synthesized sphere example of Chapter 2.8.

3.4.1 DETERMINING THE SURFACE ORIENTATION AT AN OBJECT POINT

Suppose three images $I_1(x,y)$, $I_2(x,y)$ and $I_3(x,y)$ are taken under a fixed object surface viewer geometry but with a varying light source position. Suppose the corresponding reflectance maps are $R_1(p,q)$, $R_2(p,q)$ and $R_3(p,q)$. Choose a particular image point (x_0,y_0) and suppose that the intensities at (x_0,y_0) in the three images are given by $I_1(x_0,y_0) = \alpha_1$, $I_2(x_0,y_0) = \alpha_2$ and $I_3(x_0,y_0) = \alpha_3$. One plots, in gradient space, the three contours $R_1(p,q) = \alpha_1$, $R_2(p,q) = \alpha_2$ and $R_3(p,q) = \alpha_3$. Any gradient point (p,q) lying on all three contours is a possible gradient corresponding to the image point (x_0,y_0). If there is only one such (p,q), then this (p,q) uniquely determines the local surface orientation at the object points corresponding to the given image point (x_0,y_0).

Figure 3-10 illustrates. The three images are obtained by varying the position of the light source. Let the first be $p_1 = 0.7$ and $q_1 = 0.3$, as in Chapter 2.8. Let the second and third correspond to rotations of the light source about the viewing direction of -120° and $+120^\circ$ respectively (i.e., $p_2 = -0.610$, $q_2 = 0.456$ and $p_3 = -0.090$, $q_3 = -0.756$). The phase angle g is constant in each case. Consider image point $x = 15$, $y = 20$. Here, $I_1(x,y) = 0.942$, $I_2(x,y) = 0.723$ and $I_3(x,y) = 0.505$. *Figure 3-10* shows the reflectance map contours $R_1(p,q) = 0.942$, $R_2(p,q) = 0.723$ and $R_3(p,q) = 0.505$. The point $p = 0.275$, $q = 0.367$ at which these three contours intersect determines the gradient corresponding to image point $x = 15$, $y = 20$.

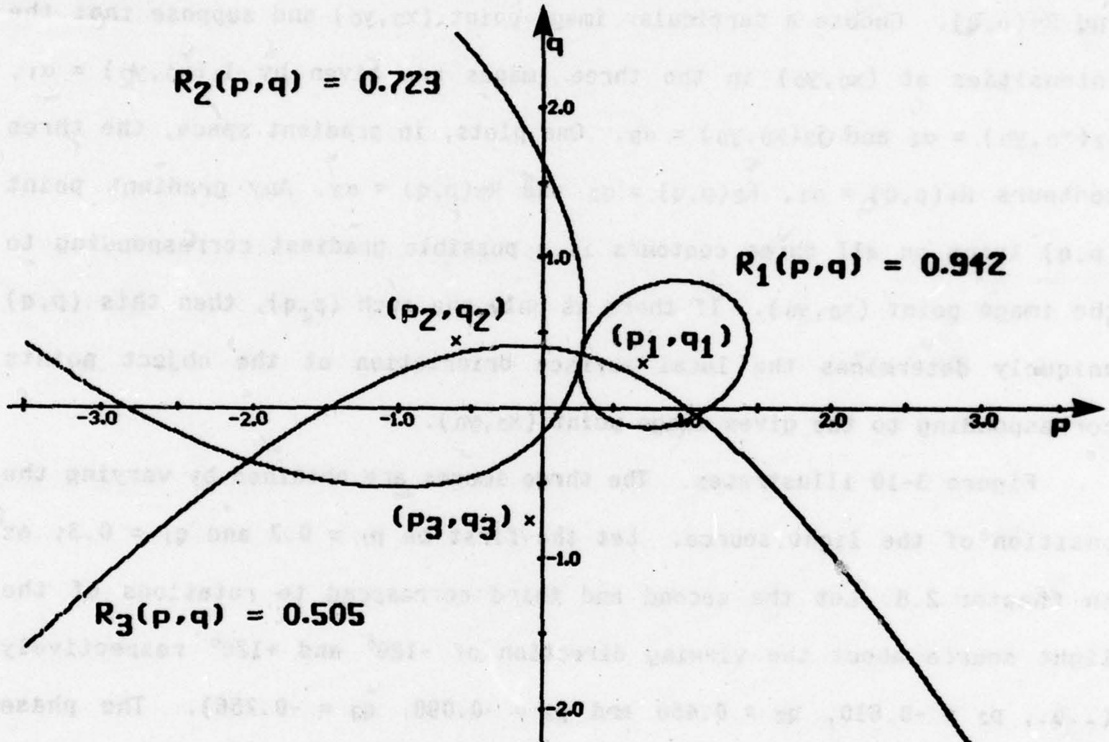


Figure 3-10 Determining the surface orientation at a given image point (x,y). Three superimposed reflectance map contours are intersected where each contour corresponds to an intensity value at (x,y) obtained from three separate images, taken under the same imaging geometry but with different light source position.

3.4.2 DETERMINING OBJECT POINTS WITH GIVEN SURFACE ORIENTATION

Suppose three images $I_1(x,y)$, $I_2(x,y)$ and $I_3(x,y)$ are taken under a fixed object surface viewer geometry but with a varying light source position. Suppose the corresponding reflectance maps are $R_1(p,q)$, $R_2(p,q)$ and $R_3(p,q)$. Choose a particular gradient point (p_0,q_0) and suppose that the reflectance values at (p_0,q_0) in the three reflectance maps are given by $R_1(p_0,q_0) = \alpha_1$, $R_2(p_0,q_0) = \alpha_2$ and $R_3(p_0,q_0) = \alpha_3$. One plots, in image space, the three contours $I_1(x,y) = \alpha_1$, $I_2(x,y) = \alpha_2$ and $I_3(x,y) = \alpha_3$. There may be zero, a finite number or an infinite number of image points (x,y) lying *on all three contours*. If there is no such (x,y), then there is no object point in view with surface orientation given by the *gradient* (p_0,q_0). If there are a finite number of such (x,y), then, barring degeneracy, each (x,y) corresponds to an object point with local surface orientation given by the gradient (p_0,q_0). If the surface is not doubly curved, then the set of such (x,y) may be infinite (i.e., the surface may have surface orientation given by gradient (p_0,q_0) over a curve *or* region in the image).

Figure 3-11 illustrates. Consider gradient point $p = 0.5$, $q = 0.5$. Here, $R_1(p,q) = 0.974$, $R_2(p,q) = 0.600$ and $R_3(p,q) = 0.375$. Figure 3-11 shows the image contours $I_1(x,y) = 0.974$, $I_2(x,y) = 0.600$ and $I_3(x,y) = 0.375$. The point $x = 24.5$, $y = 24.5$ at which these three contours intersect determines an object point whose gradient is $p = 0.5$, $q = 0.5$.

Figure 3-12 shows how to use this technique to determine a pseudo-origin (i.e., an image point corresponding to an object point with gradient $p = 0$ and $q = 0$). Here, $R_1(p,q) = R_2(p,q) = R_3(p,q) = 0.796$. Object points with surface *normal* directly facing the viewer form a unique class of points whose image intensity is invariant for rotations of the

light source about the viewing direction. The point $x = 0$, $y = 0$ at which these three contours intersect determines an object point with surface normal directly facing the viewer. This result holds even if the form of $R(p,q)$ is unknown.

3.4.3 USING PHOTOMETRIC STEREO

With any stereo technique, there is some trade-off to acknowledge. In photometric stereo, choosing a large phase angle g allows one to achieve a better "conditioning" of the gradient contours that must be intersected from each image. At the same time, a large phase angle g puts more of gradient space into the shadow region of one or more of the sources. In the example below, a four-source scheme is considered. Each source is a 90° rotation from its two neighbors with the first source again positioned at $p_S = 0.7$, $q_S = 0.3$. With this configuration, all object points within 61.7° view angle are illuminated by at least three independent sources. Further, assuming lambertian reflectance, contours in sections of gradient space illuminated by only two independent sources intersect uniquely. With this four-source scheme, photometric stereo uniquely determines the gradient corresponding to an arbitrary image point.

Figure 3-13 illustrates this four-source scheme applied to the sample sphere of Chapter 2.8. Here, all gradient points corresponding to sampled image points have been pinned down exactly. Figure 3-14 superimposes on figure 3-13 the light source positions and corresponding shadow-lines. These shadow-lines divide gradient space into nine regions, each illuminated by a different combination of the four sources.

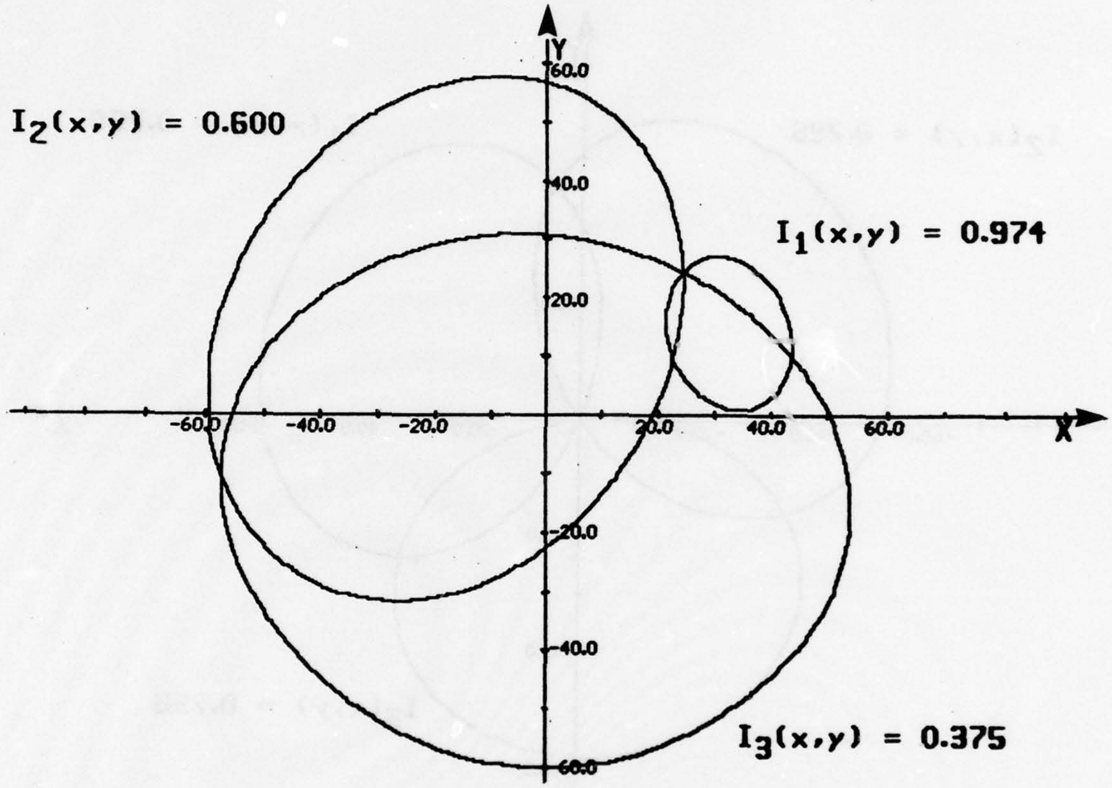


Figure 3-11 Determining image points whose surface normal is given by the gradient (p,q). Three superimposed image intensity contours are intersected where each contour corresponds to the value at (p,q) obtained from three separate reflectance maps. Each reflectance map characterizes the same imaging geometry but corresponds to a different light source position.

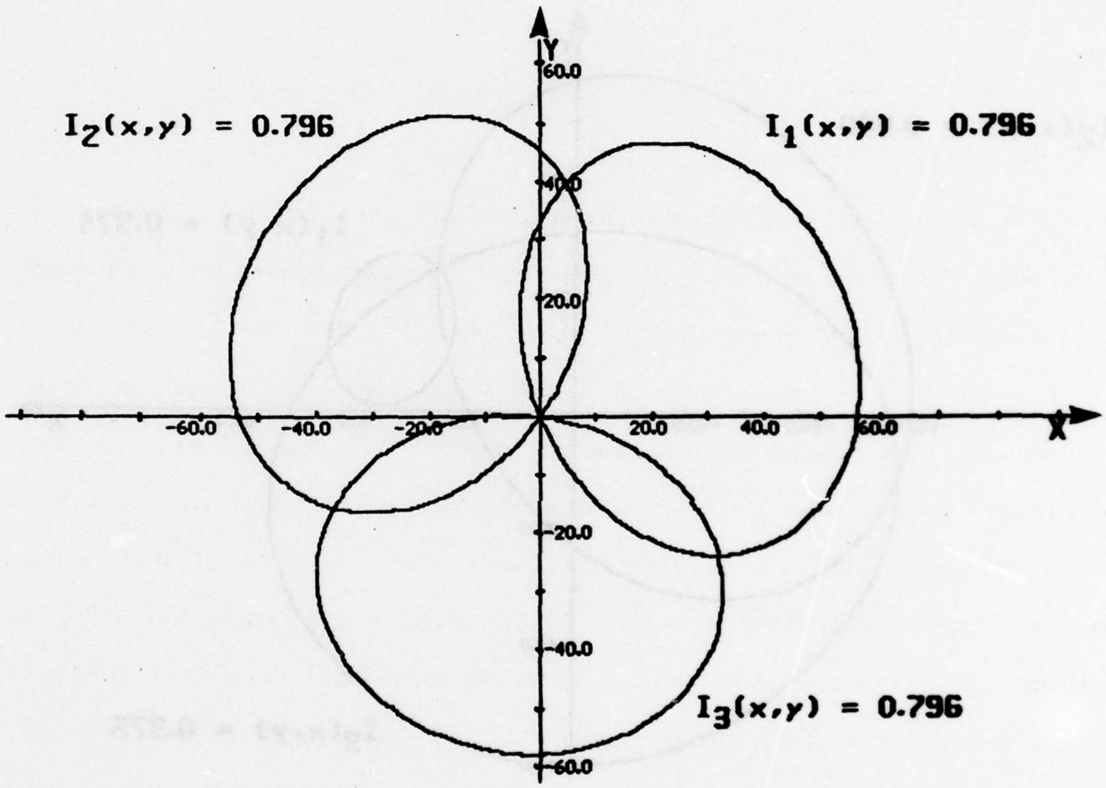


Figure 3-12 Determining image points whose surface normal directly faces the viewer. Three superimposed image intensity contours are intersected where each contour corresponds to the value at (0,0) obtained from three separate reflectance maps. Each reflectance map characterizes the same imaging geometry but corresponds to a different light source position.

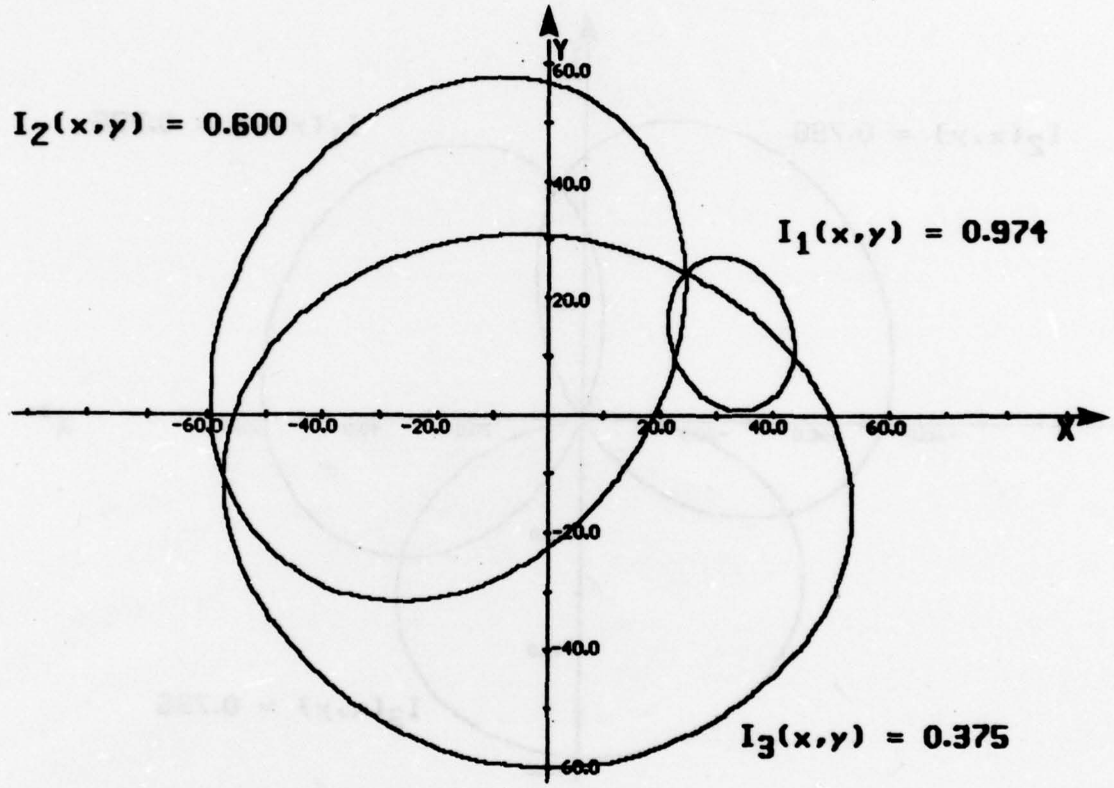


Figure 3-11 Determining image points whose surface normal is given by the gradient (p,q). Three superimposed image intensity contours are intersected where each contour corresponds to the value at (p,q) obtained from three separate reflectance maps. Each reflectance map characterizes the same imaging geometry but corresponds to a different light source position.

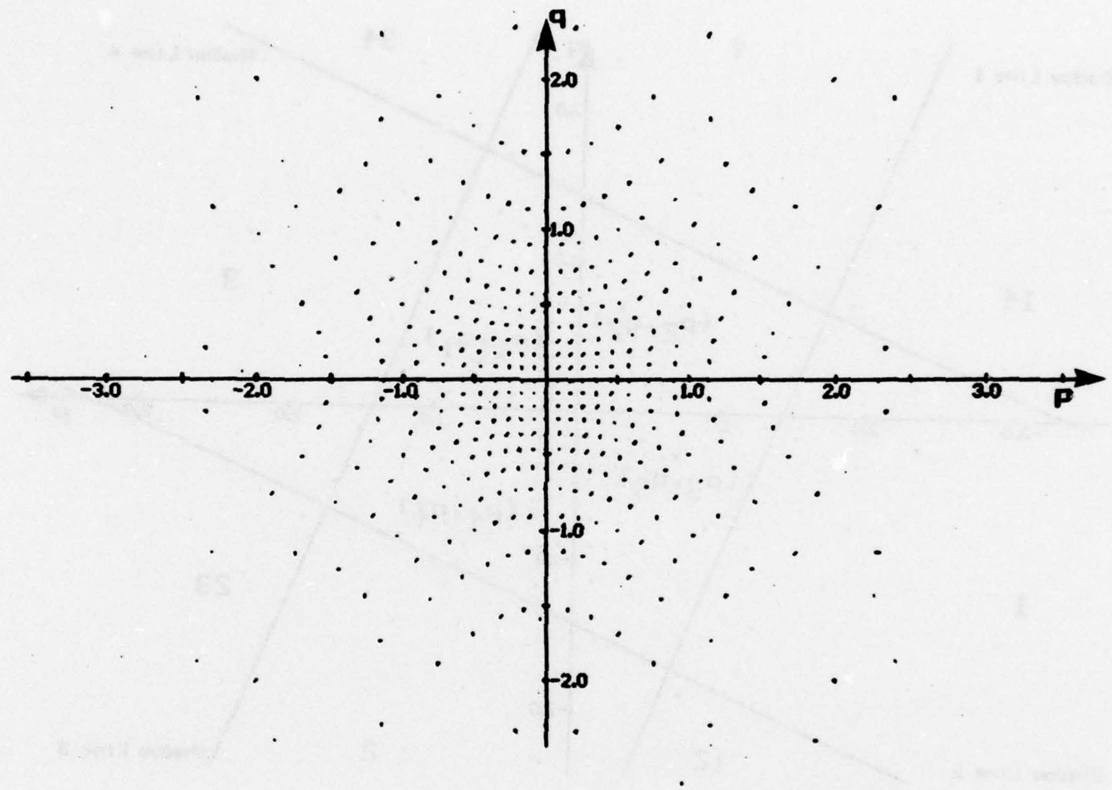


Figure 3-13 Applying photometric stereo to four synthesized images of a sphere. The surface reflectance is assumed to be lambertian. The points mark the gradients determined at each sampled image point. (Compare this figure with the single image result of Figure 2-10.)

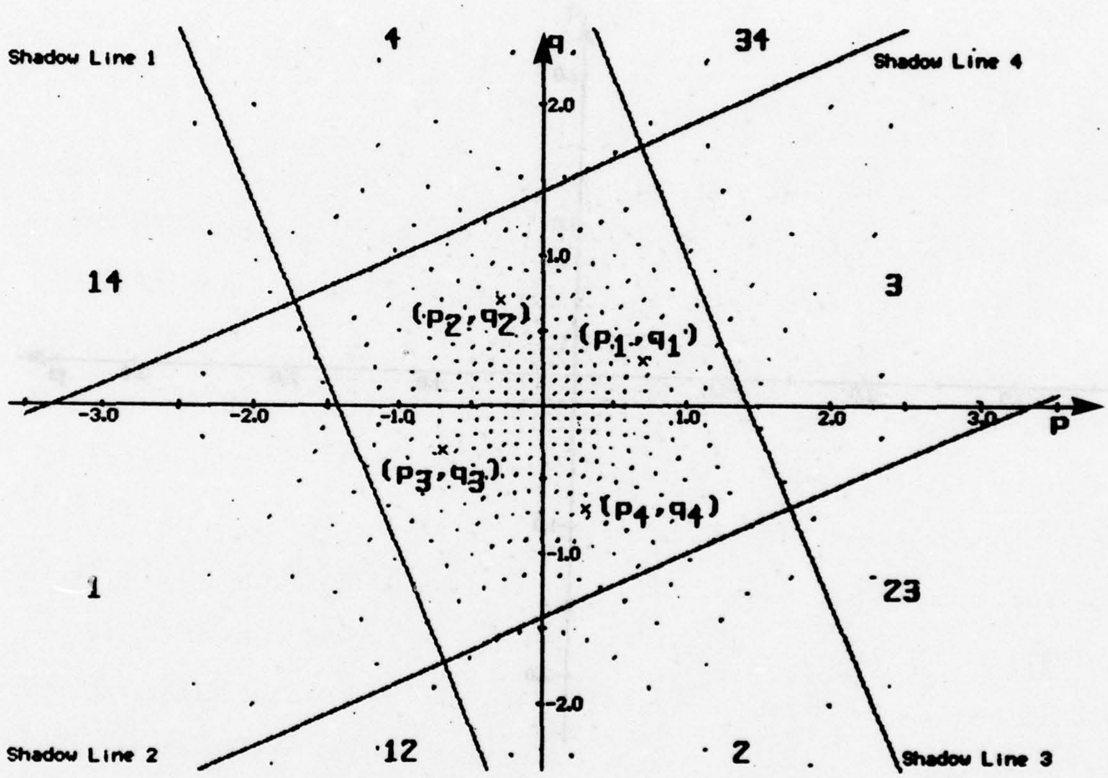


Figure 3-14 Superimposing the light source and corresponding shadow-line positions used to determine figure 3-13. This four source configuration divides gradient space into nine regions. The numbers in each region indicate which sources can not illuminate that region.

Of course, photometric stereo works well on synthesized examples. What difficulties might one expect in practice? Nonuniformities in imaging hardware, surface photometry and incident illumination, due, for example, to mutual illumination, will all lead to inaccuracies. One can gain some insight into how such inaccuracies will affect photometric stereo. Figure 3-15 plots the reflectance map contours associated with a two source lambertian configuration (spaced 0.1 units apart) when the light sources are separated by 90^{0}. Each region of Figure 3-15 corresponds to a region of equal measurement error. The configuration depicted in figure 3-15 is quite tolerant of errors for image points whose corresponding gradient lies in the third quadrant of gradient space. On the other hand, slight errors in the measurement of intensity for image points whose gradient lies in the first or second quadrant can lead to substantial errors in the solution determined for that gradient. Figure 3-16 repeats figure 3-15 but for the case when the light sources are separated by 180^{0}. Here, reasonable accuracy is achieved in the second and fourth quadrants.

This analysis is purposely quite qualitative. In any numerical solution to an overdetermined problem, which is subject to measurement error, one cannot expect to achieve an exact solution. Figure 3-15 and figure 3-16 give some indication about about how one might select which image combinations to believe and which to hold suspect. In practice, the properties of the particular reflectance map must also be taken into account.

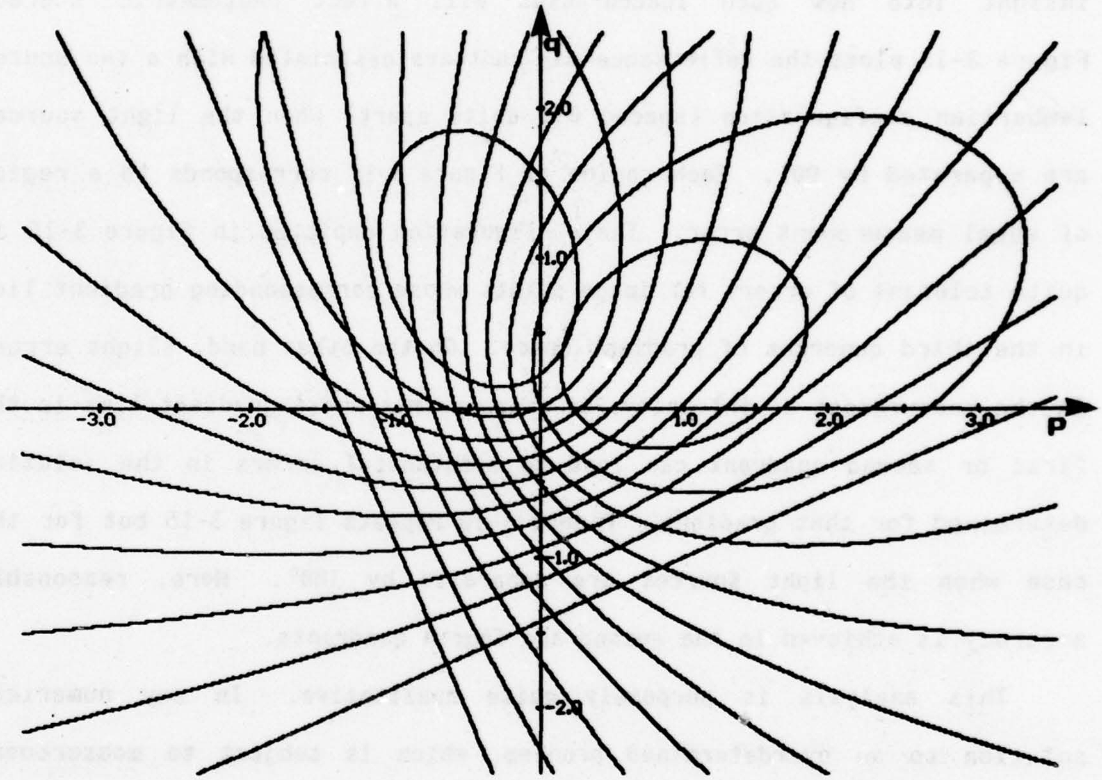


Figure 3-15 Error regions for two superimposed reflectance maps corresponding to lambertian surfaces illuminated by separate sources rotated about the viewing direction 90° from each other (g = 37.3°). Each region indicates how an error in intensity measurement determines a corresponding error in the estimation of surface orientation.

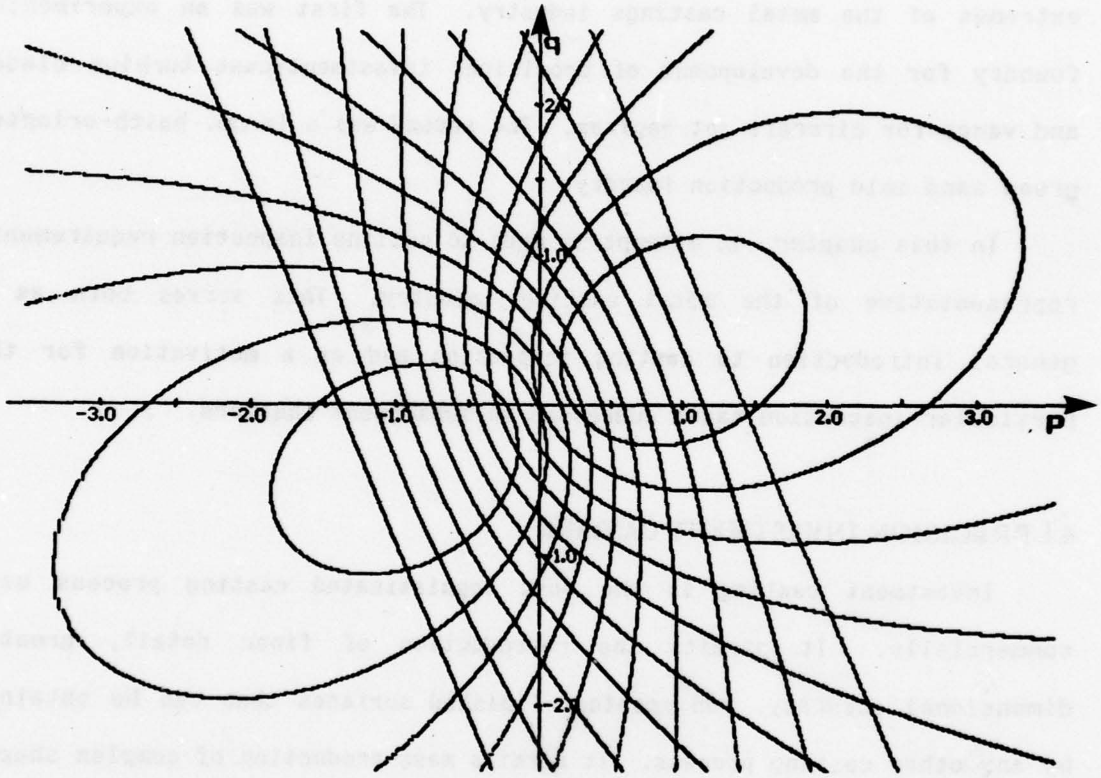


Figure 3-16 Error regions for two superimposed reflectance maps corresponding to lambertian surfaces illuminated by separate sources rotated about the viewing direction 180° from each other ($g = 37.3°$). Each region indicates how an error in intensity measurement determines a corresponding error in the estimation of surface orientation.

4. TWO CASTING APPLICATIONS

In order to understand the inspection requirements of the casting industry, it is vital to obtain actual exposure to realistic foundry environments. In the course of this work, two foundries were visited that represent two extremes of the metal castings industry. The first was an experimental foundry for the development of precision, investment-cast turbine blades and vanes for aircraft jet engines. The second was a large, batch-oriented green sand mold production foundry.

In this chapter, an attempt is made to outline inspection requirements representative of the metal casting industry. This serves both as a general introduction to casting inspection and as a motivation for the particular inspection tasks addressed in subsequent chapters.

4.1 PRECISION INVESTMENT CASTING

Investment casting is the most sophisticated casting process used commercially. It permits the reproduction of finer detail, greater dimensional accuracy, and smoother finished surfaces than can be obtained by any other casting process. It permits mass production of complex shapes that are difficult or impossible to produce by conventional casting processes or by machining. Castings can be produced that require little or no finishing for completion, thus minimizing the importance of selecting easy-to-machine metals.

Investment casting allows close control of grain size, grain orientation and other solidification conditions which results in close control of mechanical properties. The process is adaptable to almost any metal that can be melted and poured. It is also adaptable to the melting and casting of alloys that must be poured in a vacuum or under the

protection of an inert atmosphere.

4.1.1 ONE PART IN THE MAKING

The first class of parts to be discussed is that of turbine blades and vanes for aircraft jet engines. Early turbine airfoils were forged. Now, however, forging has been replaced by investment casting. The reasons for this switch are twofold:

- A gas turbine becomes more efficient as its operating temperature is raised. Alloy development for superior high-temperature performance made forging impractical.
- Sophisticated air cooling schemes have been developed to permit operation at even higher temperatures. The internal passageways necessary for cooling can be produced only by casting.

The particular turbine blades and vanes studied here are made of a special high temperature nickel-base alloy using a precision shell investment casting process. The initial tooling cost for this type of casting is substantial. Consequently, in this environment, one typically deals with relatively few part types produced at a relatively high volume or, as is the case here, with a part type for which there is no alternative manufacturing process available.

Casting is basically the process of pouring molten metal into a mold, letting the metal solidify and then separating the part from the mold. In certain casting processes, such as die casting, the mold can be used again. In other casting processes, such as green sand mold casting, the mold is destroyed during part removal. Part reproducibility is maintained in the pattern used to produce the mold.

In investment casting, both the pattern and the mold are destroyed for each part made. In this case, part reproducibility is maintained in the accurately machined metal dies used to produce each pattern. Investment casting employs a ceramic mold. The mold is produced by surrounding an expendable pattern, made of wax or plastic, with a refractory slurry. After the mold has set, the pattern is melted or burned out, creating the mold cavity. Finally, the mold is kiln hardened.

In this particular application, the inside of each turbine airfoil is hollowed out in an elaborate pattern. This pattern produces the internal passageways used for air cooling. To cast these passageways, a core (i.e., a negative image of the volume to be hollowed out) is itself produced out of a silica material. This core is fixed in the wax pattern by inserting it between the die halves prior to the injection of the pattern wax. Thus, after the wax has been burned away from the ceramic mold, the core becomes part of the mold.

Let us consider more detail. The ceramic shell mold is prepared by alternately dipping the pattern assembly (pattern and core) into slurries of ceramic powders suspended in a liquid, draining the excess, stuccoing the wetted surface with a dry refractory grain and drying the resultant coating. This process is repeated until the desired mold thickness is achieved, generally 1/4" to 1/2". The initial coating employs a slurry that is made up of particles finely ground to provide the desired surface smoothness. Subsequent coatings contain increasingly coarser refractory grains.

The mold is dried after each dipping operation to allow bonding of its individual layers. Moisture removal is regulated by control of wet and dry bulb temperatures, air flow and time. The completed mold is allowed to dry

for a minimum of 24 hours. When the mold has dried out, the wax is burned out leaving the silica core suspended in the mold cavity. Final curing of the mold is achieved by kiln hardening.

Most high-temperature superalloys are melted using a vacuum-induction furnace. Vacuum melting reduces the amount of dissolved gases in the metal. It also prevents contamination of the metal by the gaseous elements present in normal air atmospheres and actually helps purify the metal through volatilization of existing impurities. Vacuum melting reduces the number of resulting inclusions in the casting and thus improves the mechanical properties of the metal.

The molds are readied in a preheat furnace and then moved to the mold locker chamber. After this chamber has been evacuated, the interlock between it and the main furnace chamber is opened and the mold is moved into pouring position. When the specified pouring temperature has been reached, the metal is poured into the mold at a controlled rate. The mold is then withdrawn back into the mold locker chamber and the interlock closed. The mold chamber is then opened to allow removal of the filled mold.

After solidification, the ceramic mold material is broken away and the casting is cleaned by sand blasting. A first visual inspection is performed to catch obvious defects, such as those caused by a crack in the mold. After this inspection, the casting is immersed in a strong caustic solution to eat away the silica core material still embedded in the casting.

Each casting then undergoes exhaustive inspection. In an application as critical as the production of components for aircraft jet engines, part integrity is of paramount importance. In the United States, castings for

aircraft engines are 100% inspected (i.e., each casting undergoes each inspection operation). The inspection operations performed are summarized below:

- A coordinate gauging machine is used to verify the outer dimensions of the casting.
- Ultrasonic transducers are applied to selected points on the casting to verify the inner dimensions of the casting. In this case, it is to verify that the core was positioned correctly.
- The casting is X-rayed from multiple views and the resulting images are checked for indications of inclusions and other internal defects.
- Flourescent penetrant inspection is performed on all visible surfaces of the casting to check for surface cracks, tears, porosity and inclusions.
- The casting is acid-etching to verify grain structure.
- Pressure/leak testing is performed on the internal passageways of the casting to verify that none are blocked due to residual core material or due to excess metal resulting from collapsed core material.

Detailed quality standards are used for each of the above inspection operations. The basic format of these quality standards is to divide the casting into areas, each of which is judged by a portion of the standard written specifically for it. Important considerations include the absolute size of each imperfection, the total number of imperfections and the relative spacing between imperfections. This division of the casting into areas is based upon the actual service requirements of the various

subsections of the casting. Critical areas subject to high stress, for example, are inspected to tighter criteria than are areas subject to lesser degrees of stress.

It is important to realize that these quality standards cannot be implemented by simple detection devices. Detection devices help delineate imperfections. But, the mere presence of imperfections does not lead to a PASS/FAIL decision. Rather, imperfections must be interpreted in the context of the part as a whole. The ability to automate existing industry standards ultimately requires the ability to relate the imperfections detected to a higher-level representation of the object shape.

4.2 GREEN SAND MOLD CASTING

Green sand molding is the most versatile and inexpensive casting process commercially available. Typical green sand mold foundries are batch-oriented, producing many different castings at low to medium production volumes. At the particular foundry visited, approximately 40% of the castings produced were to meet the internal manufacturing demands of the parent company. The remaining 60% were cast on a contract basis for outside customers. The foundry maintains a library in excess of 100,000 patterns, any one of which can be scheduled for production. Orders range from fewer than 10 parts to as many as 250,000.

In preparing a casting for production, extensive inspection and destructive test facilities are available to eliminate systematic design flaws. In each application, however, there is an inevitable trade-off reached between design cost and production cost. A decision must be made as to whether the production quantity involved warrants further effort to modify the casting design. In a batch-oriented foundry, many of the

process variables, such as alloy composition, pouring temperature and sand composition, must be considered fixed since they cannot be modified from part to part. Thus, the most economical compromise often tolerates a high rejection rate during production. The rejection rates for castings produced at the foundry visited typically vary between 6% and 30%, depending principally on the size, shape and complexity of the part.

Green sand mold casting has the advantage that all the materials required can be recycled. The pattern is reusable. The sand used to make the mold is reusable. Should the casting be defective, the metal is also reusable. Thus, a high rate of rejection is not that expensive. The cost of manufacturing a defective part is the cost of the energy required to melt the metal and the cost in time and labor associated with the loss of production capacity.

The principal goal of inspection is to find defective castings before investing expensive machining operations on a bad part. Approximately 85% to 90% of the casting defects are sufficiently gross in nature that they are easily found by unskilled inspectors immediately following mold removal and cleaning. These human inspectors perform a cursory examination of each casting as it comes out of the sandblast machines. Ones that pass this first visual inspection are sorted into appropriate bins. The remaining ones are cataloged by an inspection foreman and then recycled to the melting furnace.

A second level of visual inspection is achieved by making it profitable for in-house machinists to find defective castings. Machinists are employed to smooth away residual traces of gating and other minor surface imperfections from the castings. These machinists are paid on a piecework basis. They are paid for each piece they grind and for each

piece they find to be defective. Thus, a few seconds of visual inspection can earn as much as several minutes of grinding. In fact, one of the major problems the foundry has is that a few unnoticed swings with a hammer can make a machinist as much money as a half an hour of "regular" work.

Castings that leave the foundry have a reasonable chance of being free of defects on all visible surfaces. No inspection, however, is provided to verify the internal integrity of the part. Such castings are not produced on a "guaranteed" basis. Defects are often found during subsequent machining operations. A potential disagreement between a foundry and its customers centers around the cost of machining defective parts. The foundry does not pay for machining operations performed after a part has left the foundry. The customer, however, will often argue that the foundry should absorb these machining costs since, after all, the part was defective.

Before presenting the second class of parts discussed, it will be useful to give a short introduction to the green sand mold casting process. Sand, combined with a suitable binder, is packed rigidly about a pattern, so that when the pattern is removed, a cavity corresponding to the shape of the pattern remains. Molten metal poured into this cavity and solidified develops a cast replica of the pattern. The sand that forms the mold cavity can be readily broken away for subsequent removal of the casting.

The materials used for pattern making differ greatly in their characteristics and therefore in the applications to which they are best suited. Patterns can be made of wood, metal or other suitable materials, such as wax, polystyrene or epoxy resin. The decision as to what material to use depends on the stage of development of the design of the casting, the expected production quantity, the dimensional accuracy required, the

size and shape of casting and the molding process to be employed.

Green sand molding is the most widely used of all sand molding processes. A green sand mold is made of sand, clay, water and other materials. "Green" means that the sand mixture remains moist. Green sand molds are not oven dried or otherwise hardened. The mold is used directly without further conditioning.

The problems associated with green sand mold casting fall into roughly two categories: those associated with metal-to-mold interactions and those associated with pouring and feeding the molten metal. Defects during production runs are minimized by careful control of sand properties, metal handling and pouring rates. Table 4-1, taken from [ASM 70], illustrates typical casting defects that result when sand properties are above or below specified limits. The foundry visited maintains an extensive laboratory for the testing and control of sand properties.

Casting Defects That Result When Sand Properties
Are Above or Below Specified Limits

Sand Property	Casting Defects
	Sand Property Above Limit
Moisture content	Blows, scabs, cuts, rough finish, hot tears, porosity, rattails, dirt, high hot strength (difficult shakeout), oxide inclusions, dimensional inaccuracy
Permeability	Poor finish, pinholes, veining, sticky sand, misruns in thin sections
Green strength	Rough finish, difficult shakeout
Green deformation	Scabs
Mold hardness	Blows, scabs, hot tears, difficult shakeout
Dry strength	Hot tears, pinholes, difficult shakeout
Hot strength	Hot cracks, difficult shakeout
Hot deformation	Dimensional inaccuracy
Combustibles content ...	Blows, pinholes
	Sand Property Below Limit
Moisture content	Drops, cuts, poor finish, dirt, broken mold edges
Permeability	Blows, penetration of metal into mold, shakeout, scabs
Green strength	Drops, scabs, cuts and washes, pinholes, dirt, veining, stickiness, dimensional inaccuracy, shrinks
Green deformation	Drops, cuts, dirt
Mold hardness	Drops, cuts, rough finish, penetration of metal into mold
Dry strength	Cuts, dirt
Hot strength	Cuts, poor finish, penetration of metal into mold
Hot deformation	Scabs, rattails, dirt, veining
Combustibles content ...	Poor finish, veining, inaccuracy, burn-on, cuts, rattails

TABLE 4-1

Metal handling and pouring, on the other hand, is more difficult to control. All operations are manual. Another recurring problem in the foundry is that metal handlers, also paid on a piecework basis, will try to squeeze more mold fills out of each bucket of molten metal than production standards call for.

Green sand mold casting is not used to produce parts whose mechanical properties are of prime concern. The internal integrity of such a casting is difficult to control. On the other hand, green sand mold casting is an economical way to produce a part, out of metal, with desired geometric properties. Defects are principally surface properties which effect these geometric properties, either functionally or aesthetically. Internal defects discovered by later machining are a mild embarrassment to the foundry but must be tolerated because there is no economical way to check for them.

Most of the defects listed in Table 4-1 manifest themselves as surface properties that could not have been the result of a legitimate casting operation. Automatic inspection in a batch-oriented foundry would require a system that could detect these surface properties independent of the particular part being viewed.

4.2.1 A SECOND PART IN THE MAKING

The particular part to be discussed is a small shuttle eye (i.e., thread guide) used in textile looms. It is one of about 50 different shuttle eyes produced. Some differ slightly in part geometry. Others differ slightly in part dimensions. It is made of gray iron and measures approximately 5 cm by 2.5 cm by 2.5 cm. The size is worth mentioning because this particular casting has a property shared by many green sand mold castings. That is, if one were only interested in making this particular casting, one would not choose a green sand mold process. It is a small part of relatively complex shape. These two factors combine to make green sand molding particularly difficult. It is difficult to achieve mold stability for a part with such small, thin sections. It is

difficult to pour the molten metal so that each mold section is filled uniformly. Rejection rates for the shuttle eye are in excess of 30%. If there were only interest in making this particular part, it would probably be die cast. However, when there is already a green sand mold facility available, a high rejection rate can be tolerated.

Of those rejected, approximately 50% are due to a cold shut defect at the boundary between the two major sections of the casting. A cold shut defect occurs when two streams of molten metal meet in the mold cavity and the temperature differential between them prevents the two streams from welding together properly. It appears as a lapping or layering on the surface of the casting. Figure 4-1 shows a shuttle eye with a cold shut defect.

Approximately 25% of the rejects are due to pinhole defects. Pinholes are small cavities on the surface of the casting caused by bubbles of entrapped gas which reach the surface during solidification. Pinholes can occur anywhere on the surface of the casting although they are predominantly found on upward surfaces of the larger sections of the casting. Figure 4-2 shows a shuttle eye with a pinhole defect.

Cold shuts and pinholes are the two most serious defects associated with this particular family of parts. The cold shut of figure 4-1 implies a potential mechanical weakness between the two sections of the casting. But the major reason for rejection due to either a cold shut or pinhole defect is aesthetic. The surface of the casting seen in figure 4-1 and figure 4-2 is to be used "as cast." No further machining is called for. Thus, those surfaces must "look right" whether or not any imperfection would actually effect the functional performance of the part.

Figure 4-1 Image of a shuttle eye with a cold shut defect at the interface of the two major sections of the casting.

Figure 4-2 Image of a shuttle eye with a pinhole defect.

No other single kind of defect accounts for more than 10% of the rejection rate. The two most predominant remaining defects are misruns (i.e., incomplete filling of sections of the mold cavity) and dirt blowouts (i.e., loose sand in the mold cavity, which, because of its wetness, explodes on contact with the molten metal). Misruns are more difficult to detect because, in order to decide that a particular section of the mold cavity has not filled completely, it is necessary to have some knowledge of what the part ought to have looked like. Dirt blowouts, on the other hand, are always quite obvious since the "explosion" of wet sand tears up a large portion of the part. A more complete catalog of typical defects in metal castings is presented in Appendix B.

4.3 AN ASIDE: THE FOUNDRY AS WORK ENVIRONMENT

Unless one has worked in a foundry, it is impossible truly to appreciate what a hostile environment it is. There are the obvious physical demands and dangers. There are also more subtle hazards. The following quote is taken from a chapter on sand molding in [ASM 70]:

> *"Silica dust from foundry operations can produce silicosis if there is sufficient exposure, in terms of time and concentration, to free crystalline silica dust of particle size below five microns. Two to twenty years (average 10 years) are required to produce a case of silicosis when dust concentrations greatly exceed the maximal allowable...*
>
> *...The oil no-bake binders for sand molds consist of three parts: a modified linseed oil, a metallic drier, and an aromatic isocyanate. Sometimes the last two create a health hazard.*
>
> *Cobalt naphthenate (6% Co), the drier most commonly used in this binder, is slightly toxic. Cobalt compounds can produce contact dermatitis and sensitization. Lead naphthenate is also used as a drier. Lead can be absorbed by the skin and cause lead poisoning, especially when mineral oil, which may induce dermatitis and irritation, is used.*
>
> *Under certain conditions of poor ventilation, the amount of aromatic isocyanate (methylene diphenyl diisocyanate, MDI) vapor liberated in shakeout operations may be a significant health hazard. Operators of shakeout devices and of overhead cranes in the area should wear an approved organic vapor respirator. Persons with known asthmatic history should not be employed on operations that release high levels of MDI vapor. To reduce decomposition of MDI, neither the sand nor the binder should be heated above 125 F. Staining of the hands by MDI can be reduced if molders use protective creams resistant to oils and solvents and wear rubber or plastic-coated canvas gloves."*

5. GRAIN SIZE ESTIMATION

Grain structure is an important process variable in many casting applications. A gas turbine, for example, becomes more efficient as its operating temperature is raised. Methods are continually being sought to enhance the high-temperature performance of turbine blades and vanes. Grain boundaries play a critical role in determining the high-temperature mechanical properties of a metal. Increases in strength generally are achieved with a reduction in ductility, and, in high-temperature alloys, one manifestation of this phenomenon is a reduction in creep strain prior to failure. Examination of failed material shows that creep cracks are initiated along grain boundaries that make a large angle to the principal stress direction [Sullivan 1976]. Consequently, both grain size and preferred grain orientation significantly alter the mechanical response of a metal. Superior performance is achieved by carefully controlling grain structure in solidification processing.

The most important factor determining grain structure is the heat flow pattern during solidification. The principal determiner of grain size is the rate of cooling during solidification. The faster the cooling the finer will be the resultant grain structure, Areas of a casting that cool very rapidly, and thus have a fine grain structure, are termed *chill zones*. The principal determiner of grain orientation is the direction of heat transfer during solidification. Grains tend to grow parallel to the direction of heat flow. Those that have an elongated aspect ratio are called *columnar grains*.

Castings which have no preferred grain orientation are called *equiaxed* castings. Castings which try for superior mechanical performance by giving a preferred grain orientation with respect to the principal stress

direction are called *directionally solidified (DS)* castings. The turbine blades and vanes considered in this work are examples of equiaxed castings. In the manufacture of these components, three qualitative grain properties are considered essential to achieve the required resistance to crack propagation:

1. Grain must be sufficiently large.
2. Grain must not have a preferred orientation.
3. There must be no sharp demarcation lines between regions of different grain size.

Considerable effort is required to produce turbine blades and vanes with grain structure satisfying these three properties. Heat flow during solidification naturally occurs out from the leading and trailing edges of the airfoil. The columnar grain which forms as a natural consequence of this heat flow is particularly undesirable since the major stress the airfoil will encounter in service occurs precisely along this direction. Turbine blades and vanes are 100% inspected to verify grain structure.

In other casting applications, a different grain structure may be desired. In aluminum castings, for example, fine grain is generally desirable. The coarseness of porosity is proportional to grain size. Consequently, porosity is finer and less harmful in fine grain aluminum castings. Also, certain mechanical properties, such as tensile strength, are usually superior for fine grain aluminum castings. Finally, for aluminum alloys, fine grain minimizes shrinkage, causing castings to be sounder.

5.1 ESTIMATING THE AVERAGE GRAIN SIZE OF METALS

The problem in nondestructive grain size estimation is to determine the average grain size of a metal structure, a three-dimensional property, from grain cross sections observed on a two-dimensional surface passing through that structure. It is important to recognize that such an estimation of grain size can never be an exact measurement. A metal structure is an aggregate of three-dimensional crystals of varying sizes and shapes. Even if all these crystals were identical, the grain cross sections, produced by a random surface of observation through such a structure, would have a distribution of areas varying from a maximum value to zero, depending upon where the surface cuts each individual crystal. Clearly, no two fields of observation can be exactly the same. In an equiaxed casting, the size and location of grains is normally completely random.

The American Society for Testing and Materials recommends three basic methods for grain size estimation as part of ASTM Designation: E 112-74 "Standard Methods for Estimating the Average Grain Size of Metals"[1]:

1. Comparison Procedures:

   Comparison procedures are used for completely recrystallized or cast materials with equiaxed grain. They involve a direct comparison of a representative field of the test specimen with an appropriate collection of standard grain size pictures. The standard which most closely matches the test specimen is selected and the corresponding grain size is recorded.

Comparison procedures are considered convenient for human inspectors and sufficiently accurate for most commercial purposes. However, experience has shown that unless the appearance of the standard grain size series reasonably well approaches that of the sample, errors may occur. Thus, a particular standard grain size series does not adapt well to different alloys or different methods of specimen preparation.

2. Planimetric Procedures:

Planimetric procedures involve counting the number of grains in a known area of the test specimen, usually a circle or rectangle. The sum of all the grains included completely within the known area plus one half the number of grains intersected by the circumference of the area gives the number of equivalent whole grains, measured at the magnification used, within the area. A simple computation, based on an assumed formal relationship between grains per unit area and grains per unit volume, is used to convert this count into an estimate of the number of grains per square millimeter. Planimetric procedures are more accurate than comparison procedures. Accuracy falls rapidly, however, when grain deviates from an equiaxed structure.

3. Intercept Procedures:

Intercept procedures involve counting the number of grains crossed by a standard pattern applied randomly to the test specimen, usually one or more straight lines or

circles. The length of the test pattern divided by count of the number of points where the test pattern is cut by a grain boundary gives the mean intercept distance, also called mean free path or Heyn intercept. It can be shown that this distance is an unbiased estimate of the mean intercept distance within the solid material in the direction, or over the range of directions, measured.

Intercept procedures are recommended particularly for structures consisting of elongated grains or for structures containing a mixture of actual grain sizes. In the absence of a specific engineering judgment to the contrary, the intercept size is always considered to be the defining grain size value. Indeed, ASTM Committee E-4 on Metallography, under whose jurisdiction the consensus standards for grain size estimation are written, adopted the following as their official position on measurement of grain size:

> *"The referee method should be the intercept method, and the defining equation for grain size number should be that presented in Methods E 112."*
>
> -- 1974 Report of Committee E-4 on Metallography
> ASTM Proceedings, Volume 75, 1975

Intercept procedures using circular test patterns automatically compensate for departures from equiaxed grain, without giving too much weight to any local portion of the field. Ambiguous intersections at the ends of linear test lines are eliminated. Circular intercept procedures are considered most suitable for grain size estimation in quality control.

In light of these recommendations, the method explored for estimating the average grain size of equiaxed turbine blades and vanes was the three circle (Abrams) procedure specified in section 11.4 of ASTM Standard E 112-74. The goal of this work is to demonstrate that a machine vision system is capable of the performance and accuracy called for by the existing industry standard. First, the Abrams procedure is outlined and then the details of the algorithm developed to implement it are presented. Finally, grain size estimation from arbitrary views of curved surfaces is considered. The flexibility required to perform this inspection is achieved by relating grain size measurements in an image to the corresponding viewer-centered representation of surface shape.

5.2 THE THREE CIRCLE (ABRAMS) PROCEDURE

The three circle (Abrams) procedure is an intercept procedure in which it is assumed that the surface of observation is planar. This assumption of planarity is implicit in the computation used to infer the three-dimensional grain size of an equiaxed casting from measurements on a two-dimensional surface of observation. The test pattern consists of three concentric, equally spaced, circles designed to cover a total circumference of 500 millimeters on the sample surface. A little mathematical figuring will show that the diameters of these circles are 26.53 mm, 53.05 mm and 79.58 mm respectively. The procedure is as follows:

1. Perform a cursory examination of the specimen to roughly estimate its grain size. Using this rough estimate, select an image magnification that will yield approximately 100 grain crossings over the 500 mm three circle test pattern.

2. Randomly select one field for measurement and apply the test pattern to that field. If the count of grain crossings for the first application is less than 70 or more than 140, discard the result and adjust the image magnification accordingly. Repeat this step until the count is in the acceptable working range.

3. Randomly select four more fields for measurement and apply the test pattern to each field. A total of five such counts is considered sufficient to compute grain size to within 1/2 a standard ASTM grain number. If additional accuracy is required, additional fields may be sampled.

The grain size, as measured above, is indicated as a count of grain *crossings over a known path length.* Such a value is inconvenient for subsequent use. Hence, this count is normally reexpressed in terms of industry standard quantities such as nominal diameter, Feret's diameter, intercept size, specific surface, grains per unit volume, or ASTM micro-size or macro-size number.

The ASTM grain number is defined as follows:

$$G = 10.0 - 2\log_2(L)$$

where L is the mean intercept distance which, for the three circle (Abrams) procedure, is given by:

$$L = \frac{\text{length of path (in millimeters)}}{\text{total number of grain crossings}}$$

If L is determined in millimeters at an image magnification of 100X, then the resulting G is called the *micro-grain number*. If L is determined in millimeters at an image magnification of 1X, then the resulting G is

called the *macro-grain number*. The program discussed below which implements the Abrams procedure has sufficient accuracy to determine the statistically correct ASTM grain number for samples in the recommended working range. The program was tested and calibrated using sample plates available from the American Society for Testing and Materials as an adjunct to standard E 112-74. When a particular sample specimen was made to fall outside the recommended working range, the program was generally robust enough to be able to suggest an appropriate change to image magnification to bring the sample within the desired range.

5.3 A PROGRAM FOR DETERMINING ASTM GRAIN NUMBER

The initial discussion presented here focuses on the problem of applying the three circle (Abrams) procedure to images of a planar surface viewed in a direction normal to that surface. The reason for this is twofold. First, the applicable grain size measure, as defined in ASTM standard E 112-74, demands that the viewed surface be planar. Second, by viewing in a direction normal to the surface, one avoids the additional mathematics required to account for the foreshortening of path length due to an oblique viewing angle.

The program discussed makes use of simple edge mask filters to determine grain boundaries crossed by a circular test pattern. Unlike other simple detection schemes, however, the program does not make decisions based upon any single measurement. Rather, grain crossings are determined by comparing the relative response over an ensemble of different mask sizes. This makes the program usable over a broader range of image magnifications and grain sizes and, at the same time, makes it less susceptible to false indications due to noise.

First, the method used to determine grain crossings along a circular path on the surface is summarized. Second, the performance of the each portion of the program is analyzed in more detail using synthesized data. Finally, an example of the program applied to an etched section of an actual turbine vane is provided.

The algorithm for determining grain crossings along a circular path on the surface can be summarized as follows:

1. DATA ACQUISITION:

   The image is sampled along a circle centered at image point (x_c, y_c) and corresponding to a surface diameter of d millimeters. The resulting intensity values are stored as a one-dimensional (circular) array called INTENSITY.

2. INITIAL DATA PROCESSING:

   Three different edge detection filters are applied to the intensity values. Each edge filter is of the form:

   $$F_n(i) = [I(i+1)+I(i+2)+...+I(i+m)] - [I(i-1)+I(i-2)+...+I(i-m)]$$

   where I(i) denotes the i^{th} element of the array INTENSITY. The results are stored as three new one-dimensional (circular) arrays, called FILTER1, FILTER2 and FILTER3 respectively. Each FILTER array corresponds to a different choice of the value m.

3. DATA COMPRESSION:

   Each array FILTERn is processed to produce a list L_n of the form:

   $$L_n = ((\text{TYPE1 INDEX1 VALUE1}) \ldots (\text{TYPEk INDEXk VALUEk}))$$

   Each element of the list L_n denotes a local extremum in the corresponding FILTERn array. TYPE is MAXIMUM or MINIMUM, INDEX is the index of the extremum in array FILTERn and VALUE is the height of the extremum. The three lists L_1, L_2 and L_3 are then merged into a single

list

$$L = ((\text{TYPE1 INDEX1}) \ (\text{TYPE2 INDEX2}) \ \ldots \ (\text{TYPEm INDEXm}))$$

Again, TYPE is MAXIMUM or MINIMUM and INDEX is the index of the extremum. L contains an element for each local extremum found in either L_1, L_2 or L_3.

4. SYNTACTIC ANALYSIS:

The list L, together with the lists L_1, L_2 and L_3 are passed off to a syntactic parser. Basically, the job of the parser is to look at each extremum in L and decide whether it corresponds to an actual grain crossing. The parser consists of a collection of parsing routines which may accept a given extremum as a grain crossing. Each parsing routine is given initial matched pointers into L_1, L_2 and L_3. It can then look both forward and backward from those pointers using any or all of the lists L_1, L_2 or L_3 to compare the responses measured by the FILTER1, FILTER2 and FILTER3 arrays. If the responses are consistent with the type of edge being parsed, then the peak is accepted as a grain crossing.

Data acquisition is straightforward when the surface of observation is planar and is viewed in a direction normal to the surface. Circles on the object surface will project to circles in the image. To compute the ASTM grain number, it is only necessary to calibrate the imaging hardware with respect to magnification. Unfortunately, the slow-scan vidicon camera (Spatial Data System 108) used to digitize images for this work has a different spatial resolution in the image x-direction than in the image y-direction. Thus, two calibration measurements are performed. One determines A, the number of picture elements per millimeter in the x-direction and the other determines B, the number of picture elements per

millimeter in the y-direction. For a particular imaging situation, this calibration can either be done manually or automatically using an appropriately scaled "ruler" in the field of view. Thus, in order to scan a circle on the surface centered at image point (x_c, y_c) and of diameter d millimeters, one actually scans the image ellipse described parametrically by:

$$x(\theta) = x_c + A\, r \cos(\theta)$$
$$y(\theta) = y_c + B\, r \sin(\theta)$$

where $r = d/2$, the radius of the circle in millimeters and $0 \le \theta < 2\pi$.

The raw intensity values obtained with the vidicon camera are typically quite noisy and hence difficult to interpret directly. Simple edge mask filters enhance effects due to edges and suppress effects due to noise. [Shirai 75] used the same filter technique in his program for tracking edges of polyhedra. [Marr 76] uses a similar type of edge mask, extended in the second dimension, as one of the two basic operations in the computation of the primal sketch.

The contribution of the present work is not so much in the use of the simple edge mask filters but in the use of multiple filters of varying support widths. For any particular choice of m, the filter array computed by:

$$F_n(i) = [I(i+1)+I(i+2)+\ldots+I(i+m)] - [I(i-1)+I(i-2)+\ldots+I(i-m)]$$

forces an inevitable trade-off between noise immunity and resolution. For small m, there will be a greater number of false peaks due to noise. For large m, there will be a loss of resolution as closely spaced edges become smeared together, especially in situations where adjacent edges have intensity gradients with the same sign.

For a given image, it is usually possible to hand-pick a choice for m that works reasonably well. In the first version of the grain size estimation program, a single FILTER array was computed corresponding to a single choice of m. Unfortunately, no single value of m proved satisfactory over the range of grain sizes called for in ASTM standard E 112-74. With three filter arrays, corresponding to three choices of m, superior noise immunity and resolution is achieved. The existence of a peak in one of the three filter arrays is used to hypothesize the presence of an edge. There may not be matching peaks in the other two filters arrays, but if the values of the other two filter arrays at the position of the peak are as predicted by the hypothesized edge, then the existence of that edge can be reliably asserted.

Figure 5-1 illustrates edge mask preprocessing applied to a single idealized grain crossing. Figure 5-1(a) shows the intensity profile corresponding to a step in intensity. Figure 5-1(b), figure 5-1(c) and figure 5-1(d) show, respectively, the filter profiles corresponding to choices of m_1, m_2 and m_3 where the ratio $m_1:m_2:m_3$ is 1:2:3. The step in intensity produces a simple peak in each filter profile. The heights and widths of the corresponding peaks are also in the ratio $m_1:m_2:m_3$.

Figure 5-2 illustrates edge mask preprocessing applied to the same idealized grain crossing of figure 5-1. In this case, however, each intensity value in figure 5-2(a) has been perturbed by adding a white noise of amplitude h, where h is the height of the original step in intensity. Figure 5-2(b), figure 5-2(c) and figure 5-2(d) show, respectively, the filter profiles corresponding to choices of m_1, m_2 and m_3 where the ratio $m_1:m_2:m_3$ is again 1:2:3. All three filter profiles have successfully isolated the edge due to the step in intensity from the effects due to

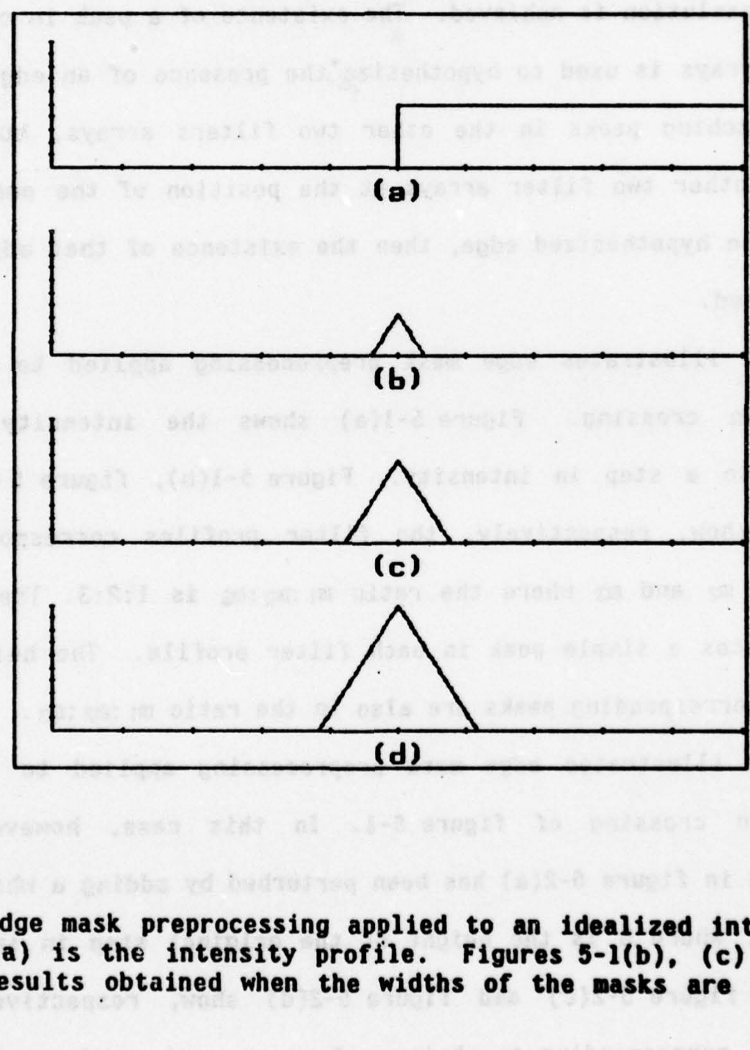


Figure 5-1 Edge mask preprocessing applied to an idealized intensity step. Figure 5-1(a) is the intensity profile. Figures 5-1(b), (c) and (d) are the mask results obtained when the widths of the masks are in the ratio 1:2:3.

noise. The height of each peak in the filter profiles preserves the approximate ratio 1:2:3. So far, there seems to be no advantage to having computed three separate filter arrays.

Figure 5-3 illustrates the superior noise immunity that can be obtained using three separate filter arrays. Figure 5-3(a) again corresponds to the idealized grain crossing of figure 5-1 perturbed by white noise. In figure 5-3(a), however, the amplitude of the noise was increased to 1.5h, where h is again the height of the original step in intensity. Figure 5-3(b), figure 5-3(c) and figure 5-3(d) show, respectively, the filter profiles computed using the same choices for m_1, m_2 and m_3 used in figure 5-2. There is a peak in the FILTER1 array corresponding to the underlying step in intensity. This peak, however, hardly "stands out" from the other peaks due to noise found in FILTER1. Note, however, that the height of the peaks in the three filter profiles corresponding to the underlying step in intensity preserve the approximate ratio 1:2:3. Other false peaks in FILTER1 do not preserve this ratio across the three filter arrays. Thus, they can be reliably rejected.

One might argue that the value chosen for m_1 is simply too small for the noise levels present in the example of figure 5-3. The next two figures illustrate that such a small value for m_1 is necessary to resolve closely spaced edges. Figure 5-4 illustrates an idealized example corresponding to two closely spaced grain crossings such that the step in intensity across each boundary has the same sign. Figure 5-4(a) shows the intensity profile. Figure 5-4(b), figure 5-4(c) and figure 5-4(d) show, respectively, the filter profiles corresponding to the same choices of m_1, m_2 and m_3 used above. FILTER1 clearly resolves the two steps in intensity. FILTER2 has begun to smear the two peaks together. FILTER3 has smeared the

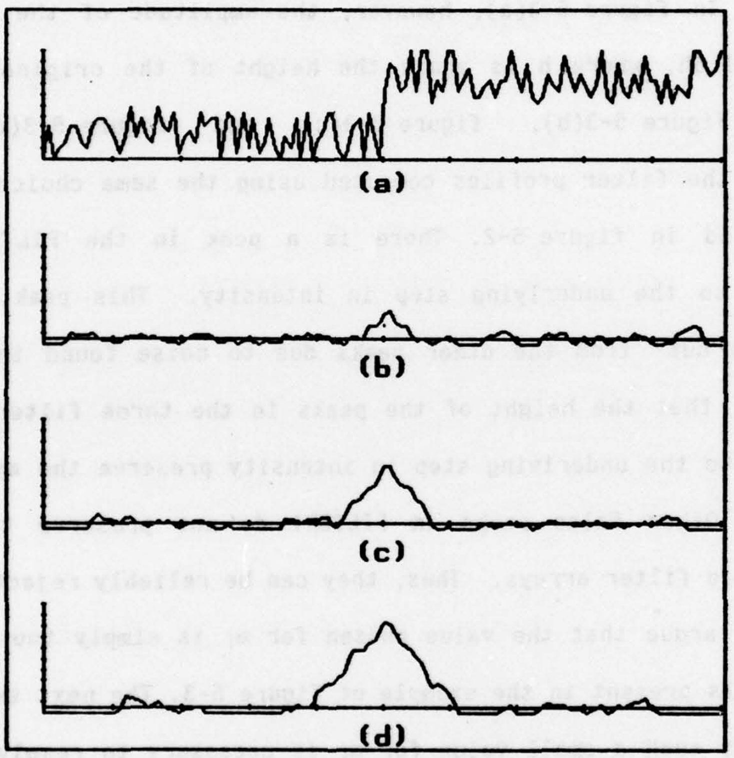


Figure 5-2 Edge mask preprocessing applied to a noisey intensity step. Figure 5-2(a) is the intensity profile. Figures 5-2(b), (c) and (d) are the mask results obtained when the widths of the masks are in the ratio 1:2:3. Each mask has isolated the step and the approximate ratio of the peak heights is preserved.

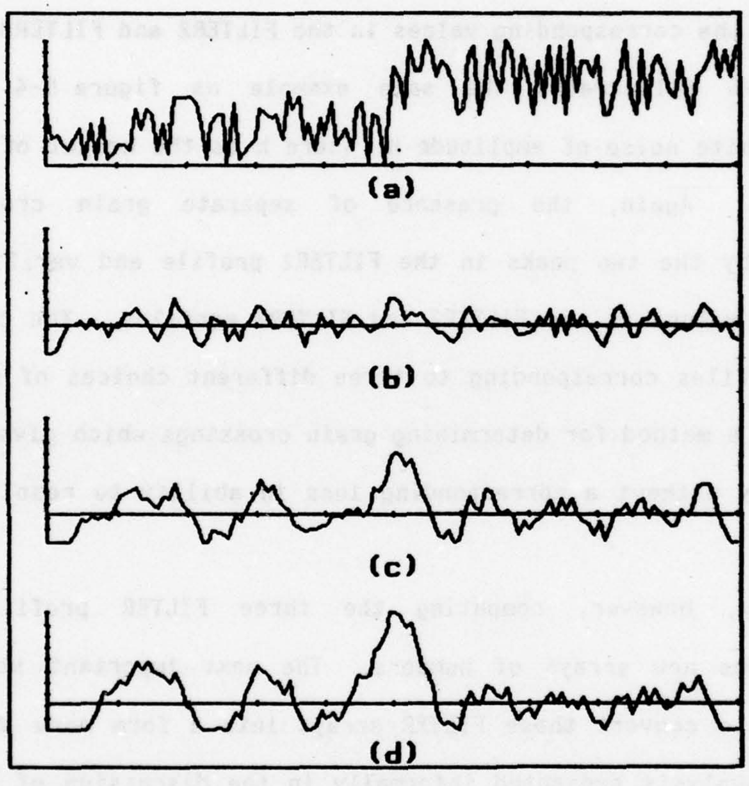


Figure 5-3 Edge mask preprocessing applied to a noisier intensity step. Figure 5-3(a) is the intensity profile. Figures 5-3(b), (c) and (d) are the mask results obtained when the widths of the masks are in the ratio 1:2:3. In figure 5-3(b), the peak due to the step is no higher than peaks due to noise. Nevertheless, the approximate ratio of the peak heights at that point confirms the presence of the step.

two peaks together to the point that it is no longer possible to resolve them. Nevertheless, the values in the three filter profiles, at each peak position in the *FILTER1 profile*, preserve the approximate ratio 1:2:3. Thus, the presence of two separate grain crossings, as hypothesized by the two peaks in the FILTER1 profile, can be reliably asserted, not due to the presence of corresponding peaks in the FILTER2 and FILTER3 profiles but rather due to the corresponding values in the FILTER2 and FILTER3 profiles.

Figure 5-5 illustrates the same example as figure 5-4 with the addition of white noise of amplitude h, where h is the height of each step in intensity. Again, the presence of separate grain crossings is hypothesized by the two peaks in the FILTER1 profile and verified by the corresponding values in the FILTER2 and FILTER3 profiles. The computation of filter profiles corresponding to three different choices of m provides the basis for a method for determining grain crossings which gives superior noise immunity without a corresponding loss in ability to resolve closely spaced edges.

Initially, however, computing the three FILTER profiles simply generates three new arrays of numbers. The next important step in the algorithm is to convert these FILTER arrays into a form more amenable to the kind of analysis presented informally in the discussion of figure 5-1 through figure 5-5 above.

The computation of each list L_n, where

$$L_n = ((\mathrm{TYPE1\ INDEX1\ VALUE1}) \ldots (\mathrm{TYPEk\ INDEXk\ VALUEk}))$$

serves two functions. First, it represents each FILTER array as a set of symbolic assertions which serves as the appropriate input to the syntactic parser. Second, it can also be properly viewed as a data compression process. The list L_n carries information only about the position, height

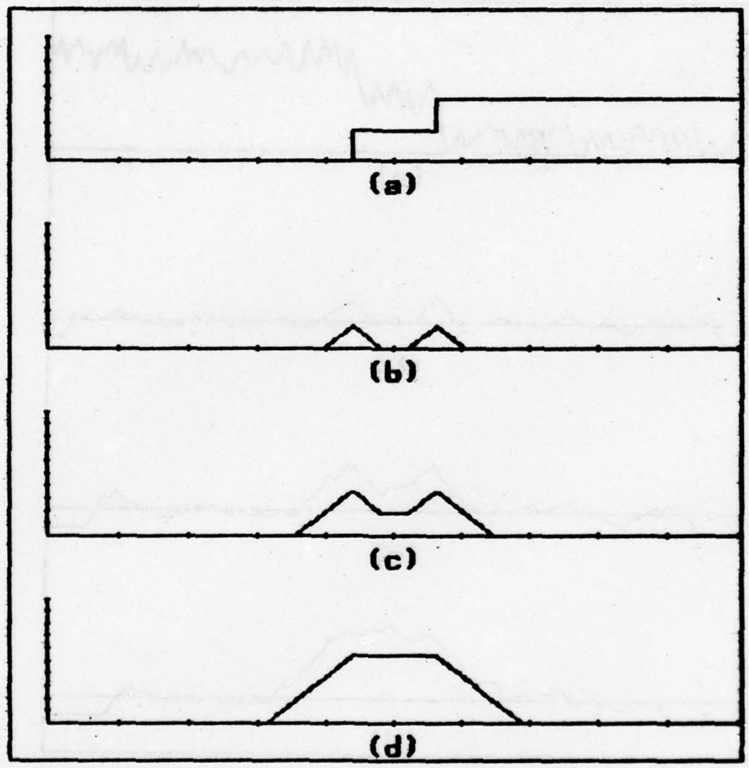


Figure 5-4 Edge mask preprocessing applied to two closely spaced idealized intensity steps (of the same sign). Figure 5-4(a) is the intensity profile. Figures 5-4(b), (c) and (d) are the mask results obtained when the widths of the masks are in the ratio 1:2:3.

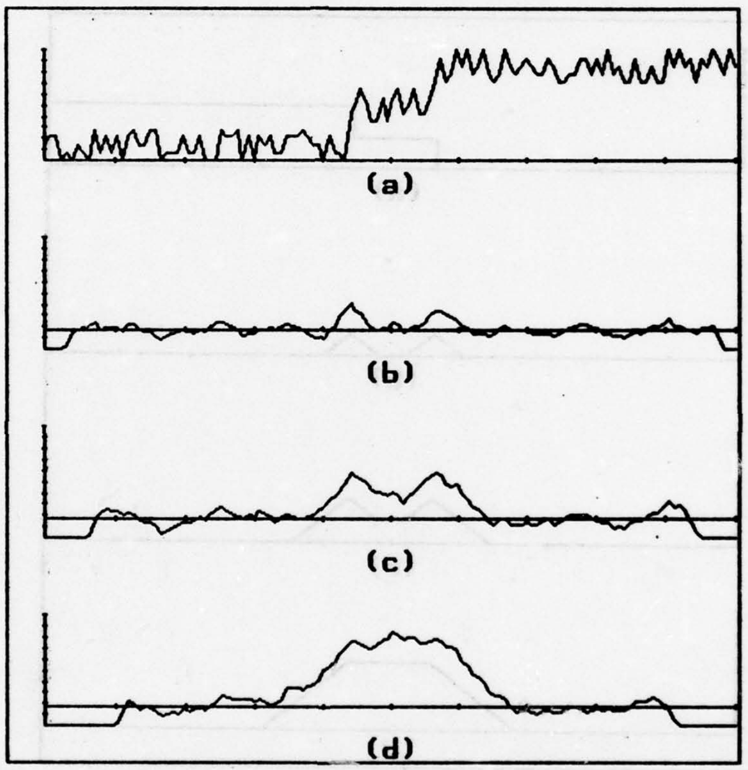


Figure 5-5 Edge mask preprocessing applied to two noisy closely spaced intensity steps (of the same sign). Figure 5-5(a) is the intensity profile. Figures 5-5(b), (c) and (d) are the mask results obtained when the widths of the masks are in the ratio 1:2:3.

and type of each local extremum in the corresponding FILTERn. In fact, one can throw away the INTENSITY and FILTER arrays at this point. It is true that subsequent analysis will require values for $F_n(i)$ where i is not a local extremum of the particular F_n, but these values can be successfully approximated as a linear interpolation between the values at the extrema neighboring i. Saying the same thing again, each L_n used in the subsequent syntactic analysis corresponds to a "stylized" $F_n(i)$ for which neighboring local extrema are joined by straight lines. In the current implementation, the FILTER arrays are available so that it is most convenient to refer back to the original arrays as required. A scheme based on linear interpolation between peaks was also tried with no change in accuracy.

The three lists L_1, L_2 and L_3 are also merged into the single list

$$L = ((TYPE_1\ INDEX_1)\ (TYPE_2\ INDEX_2)\ \ldots\ (TYPE_i\ INDEX_i))$$

The list L collects, in one place, all the peaks found in the arrays FILTER1, FILTER2 and FILTER3. This list feeds the parser candidate positions for possible grain crossings. The lists L_1, L_2 and L_3 are available to each individual parsing routine for examining particular features within each filter profile and for comparing responses across filter profiles. The process of parsing consists of passing successive peaks from L off to the individual parsing routines each of which defines a class of edge that will be accepted by the parser. When invoked, each individual parsing routine is also passed pointers to the closest matching peak in each of the lists L_1, L_2 and L_3. These individual routines may look forward or backward in any or all of the lists L_1, L_2 and L_3 in an effort to parse a candidate peak. A peak is rejected as an edge if it is rejected by all of the parsing routines present in the system. A peak is accepted as an edge when it is accepted by one of the individual parsing routines

present in the system. If a peak is accepted by one of the individual parsing routines, it is the responsibility of that routine to position the input pointer correctly past any adjacent peaks that have already been accepted as part of the edge. Parsing is complete when there are no more candidate peaks to consider.

The simplest example of a parsing routine is STEP-UP which accepts grain crossings corresponding to a simple upward step in intensity like those illustrated in figure 5-1 *through figure* 5-5. STEP-UP is invoked whenever a peak of type MAXIMUM occurs in the list L. STEP-UP, like all the individual parsing routines, is passed the four arguments: i, i1, i2 and i3 where i is the index of the candidate peak of type MAXIMUM from L and i1, i2 and i3 are, respectively, the indices in FILTER1, FILTER2 and FILTER3 of the peak nearest to i of type MAXIMUM. STEP-UP accepts a grain crossing at index i if:

(1) $F_1(i) > (\mathrm{THRESHOLD} * m_1)$

(2) $F_1(i)/m_1$, $F_2(i)/m_2$ and $F_3(i)/m_3$ have "about the same value".

STEP-UP is so simple that it does not make use of any forward or backward analysis of the three lists L_1, L_2 and L_3. Another parsing routine, PULSE-UP, discussed below, does make use of both forward and backward analysis. THRESHOLD is a global threshold which establishes the minimum step size to be considered significant. Alas, it was not possible to rid the program entirely of thresholding operations. The value of THRESHOLD, however, is not critical. The actual value used in the *program* examples is THRESHOLD = 3 which is slightly below the average noise level of the imaging hardware, measured to be about 5 gray levels. The presence of a threshold is required to rule out spurious combinations of peaks, at the

level of noise, which happen to satisfy (2) above. Finally, a set of numbers $\{v_i\}$ is said to have "about the same value" if the absolute value of the difference between any two of them is less than one half the absolute value of the one with largest magnitude. This definition corresponds to a rather ad hoc definition of qualitative equality. A more rigorous definition would have to take into account the signal to noise properties of the imaging hardware. Again, however, the particular definition of "about the same value" is not critical. In general, for choices of m_1, m_2 and m_3 in the ratio 1:2:3, there is a good separation between peak heights of "about the ratio 1:2:3" and peak heights corresponding to noise. The actual boundary chosen was not of critical importance.

The above parsing routine STEP-UP, together with a similarly defined routine STEP-DOWN, is sufficient to perform the syntactic analysis required to apply the three circle (Abrams) procedure to alloy specimens that can be prepared using a *contrast etch* technique. In a contrast etch, the cross section of each grain is etched so that surface reflectance becomes a function of grain orientation. Thus, individual grains appear as regions of homogeneous intensity. Grain boundaries appear as contrast boundaries between adjacent regions. Other alloy specimens are more appropriately prepared using a *flat etch* technique. In a flat edge, the cross section of each grain is not affected. Rather, the boundaries themselves are etched. Regions within the confines of a grain boundary are unaffected while the grain boundaries themselves appear as lines on the specimen.

The turbine blade and vane specimens used in this work were prepared using a contrast etch technique. On the other hand, the standard plates obtained from the American Society for Testing and Materials correspond to

samples prepared using a flat etch technique. Thus, additional parsing routines were also included to handle grain crossings which appear as lines in the image. Figure 5-6 illustrates edge mask preprocessing applied to a single idealized grain crossing from a specimen prepared using a flat etch technique. Figure 5-6(a) shows the intensity profile corresponding to a narrow pulse in intensity. Figure 5-6(b), figure 5-6(c) and figure 5-6(d) show, respectively, the filter profiles corresponding to choices of m_1, m_2 and m_3 where the ratio $m_1:m_2:m_3$ is once again 1:2:3. The pulse in intensity produces a doublet in each filter profile. The positive peak corresponds to the leading edge of the pulse and the negative peak corresponds to the trailing edge of the pulse. Note that if the width, w, of the pulse is small compared to m_1, m_2 and m_3, then the height of each positive and negative peak is constant in each of the FILTER1, FILTER2 and FILTER3 profiles. It is, in fact, equal to the area under the pulse. Note, also, that the transition from positive peak to negative peak occurs over the width w, again independent of which filter array is considered.

Figure 5-7 illustrates edge mask preprocessing applied to the same idealized grain crossing of figure 5-6. In this case, however, each intensity value in figure 5-7(a) has once again been perturbed by adding white noise of amplitude 0.5h, where h is the height of the pulse in intensity. Figure 5-7(b), figure 5-7(c) and figure 5-7(d) show, respectively, the filter profiles corresponding to the same choices of m_1, m_2 and m_3 used above. The same qualitative observations hold. The pulse in intensity produces a doublet in each filter profile. The magnitude of each positive and negative peak is independent of the choice of m_1, m_2 and m_3. The width of the transition from positive peak to negative peak in each filter array approximates the width of the original pulse in intensity.

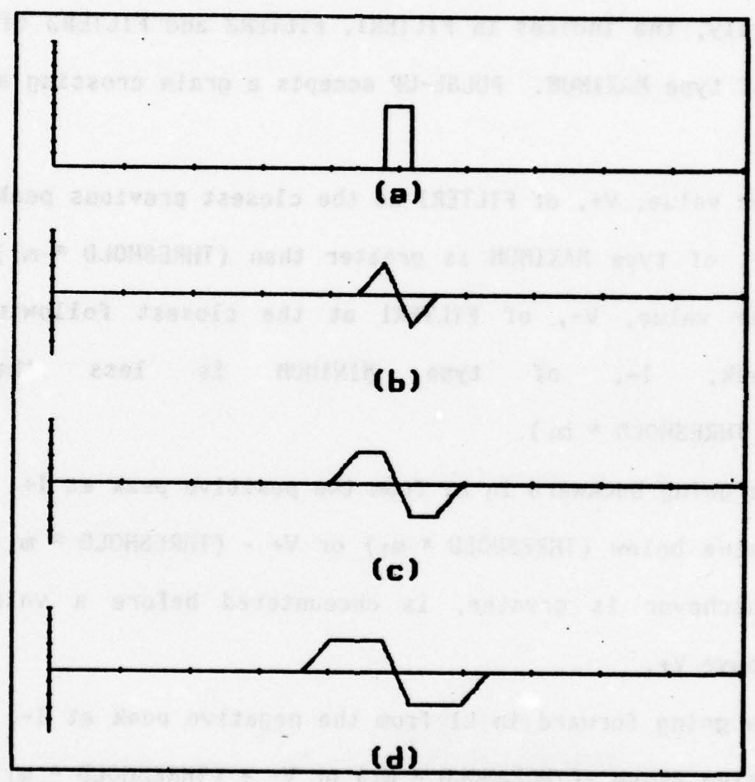


Figure 5-6 Edge mask preprocessing applied to an idealized pulse (two closely spaced intensity steps of opposite sign). Figure 5-6(a) is the intensity profile. Figures 5-6(b), (c) and (d) are the mask results obtained when the widths of the masks are in the ratio 1:2:3.

These observations form the basis for the parsing routine PULSE-UP used to accept grain crossings from a specimen prepared using a flat etch technique.

PULSE-UP is also invoked whenever a peak of type MAXIMUM occurs in the list L. PULSE-UP is passed the four arguments: i, i1, i2 and i3 where i is the index of the candidate peak of type MAXIMUM from L and i1, i2 and i3 are, respectively, the indices in FILTER1, FILTER2 and FILTER3 of the peak nearest to i of type MAXIMUM. PULSE-UP accepts a grain crossing at index i if:

(1) The value, V+, of FILTER1 at the closest previous peak, I+, of type MAXIMUM is greater than (THRESHOLD * m_1).

(2) The value, V-, of FILTER1 at the closest following peak, I-, of type MINIMUM is less than -(THRESHOLD * m_1).

(3) In going backward in L1 from the positive peak at I+, a value below (THRESHOLD * m_1) or V+ - (THRESHOLD * m_1), whichever is greater, is encountered before a value above V+.

(4) In going forward in L1 from the negative peak at I-, a value above -(THRESHOLD * m_1) or V- + (THRESHOLD * m_1), whichever is less, is encountered before a value below V-.

(5) Let the distance between the MAXIMUM peak at I+ and the MINIMUM peak at I- be w. Then:

a) w is less than m_3.

b) F_1(I+)/min{w,m_1}, F_2(I+)/min{w,m_2} and F_3(I+)/w have "about the same value".

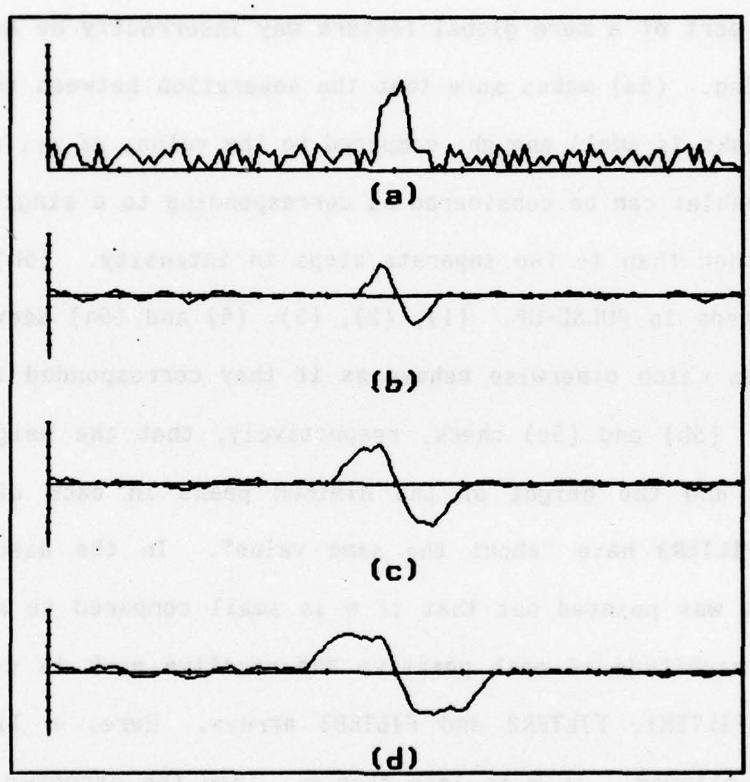


Figure 5-7 Edge mask preprocessing applied to a noisy pulse (two closely spaced noisy steps of opposite sign). Figure 5-7(a) is the intensity profile. Figures 5-7(b), (c) and (d) are the mask results obtained when the widths of the masks are in the ratio 1:2:3.

c) $F_1(I-)/\min\{w,m_1\}$, $F_2(I-)/\min\{w,m_2\}$ and $F_3(I-)/w$ have "about the same value".

THRESHOLD is the same global threshold referred to in the definition of STEP-UP. (1) and (2) rule out spurious doublets occurring at the level of noise. (3) and (4) make sure that the doublet in FILTER1 has well defined leading and trailing edges. Otherwise, a local zero crossing which actually is a part of a more global feature may incorrectly be accepted as a grain crossing. (5a) makes sure that the separation between the MAXIMUM and MINIMUM peaks is small enough, compared to the values of m_1, m_2 and m_3, so that the doublet can be considered as corresponding to a single pulse in intensity, rather than to two separate steps in intensity. (5b) and (5c) are the key steps in PULSE-UP. (1), (2), (3), (4) and (5a) serve only to eliminate cases which otherwise behave as if they corresponded to a pulse in intensity. (5b) and (5c) check, respectively, that the height of the MAXIMUM peaks and the height of the MINIMUM peaks in each of FILTER1, FILTER2 and FILTER3 have "about the same value". In the discussion of figure 5-6, it was pointed out that if w is small compared to m_1, m_2 and m_3, then the magnitude of each positive and negative peak is constant in each of the FILTER1, FILTER2 and FILTER3 arrays. Here, a little more generality is allowed. If w is less than m_i, then the expected height of each peak in FILTERi is h * w. On the other hand, if w is greater than m_i, then a pulse of width w looks to FILTERi like two separate opposite steps in intensity and the expected height of each peak in FILTERi is h * m_i.

The above parsing routine PULSE-UP, together with a similarly defined routine PULSE-DOWN, is sufficient to perform the syntactic analysis required to apply the three circle (Abrams) procedure to alloy specimens that are prepared using a flat edge technique. PULSE-UP and PULSE-DOWN are

also useful for alloy specimens prepared using a contrast etch. The routines PULSE-UP and PULSE-DOWN give the program superior ability to resolve closely spaced intensity steps in situations where the adjacent steps have intensity gradients with the opposite sign. In the case of samples prepared using a contrast etch, a peak accepted by PULSE-UP or PULSE-DOWN corresponds to two separate grain crossings.

A real example is now presented. Figure 5-8(a) shows a macro-etched section of a turbine airfoil. The airfoil has been prepared using a contrast etch. Superimposed upon the grain pattern is the three circle test pattern used in the analysis. For illustrative purposes, the test pattern is given as three concentric circles in the image rather than as the corrected three ellipses. Crosses marked along the three circles indicate positions where grain crossings were detected by the program. Figure 5-8(b) illustrates the results obtained when the program analyzed the middle circle of figure 5-8(a).

Figure 5-8(b) consists of five graphs. In each case, the abscissa corresponds to angular position (0 to 2π radians) along the test circle. The first graph shows the intensity values obtained with the vidicon camera. The next three graphs show the filter values computed for case $m_1 = 5$, $m_2 = 10$ and $m_3 = 15$ respectively. The fifth graph simply marks, for reference purposes, the points at which grain crossings were detected by the parsing routines discussed above.

A good example of how the use of three different choices of m helps can be seen in the second and third grain crossings of figure 5-8(b), counting from the left. The FILTER1 graph does have two negative peaks at those points but their magnitude is not much above that of the noise. The FILTER2 graph shows two distinct negative peaks at the points marked. The

FILTER3 graph, on the other hand, has smeared the two peaks together in such a way that no program could reasonably be expected to disambiguate them. Nevertheless, the relative magnitudes of the matched values in the three graphs are consistent with the two crossing interpretation indicated.

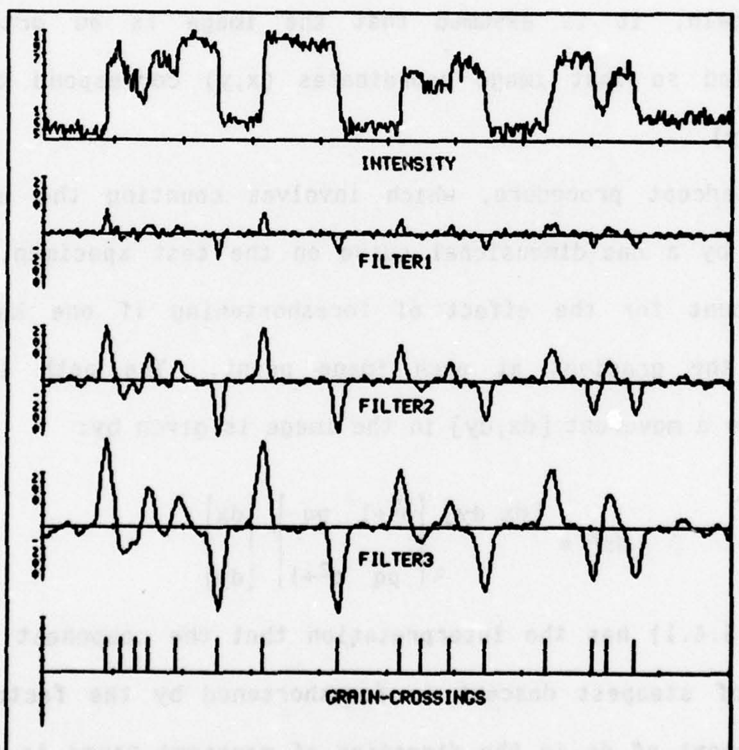
5.4 DISCUSSION

Flexibility as well as accuracy must be regarded as keys to success with any automatic inspection system. This chapter describes an important step forward in the domain of castings inspection. First, it has been demonstrated that an automatic inspection system is capable of the accuracy required to implement the existing industry standard for grain size estimation. A certain amount of flexibility has also been demonstrated in the program's ability to maintain high noise immunity without a corresponding loss in resolution.

Further flexibility requires that the measurements made in an image be related to object relief. This can be done if the corresponding viewer-centered representation of surface shape is available. A word of caution is in order, however. This discussion will describe how to make the measurements required by ASTM standard E 112-74 on curved surfaces without establishing the validity of those measurements. The goal here is not to develop a statistical theory for estimating the three-dimensional grain size of metals from arbitrary two-dimensional surfaces of observation. ASTM standard E 112-74 develops the appropriate theory for planar surfaces of observation, and it will be left at that.

Suppose one applies the three circle (Abrams) procedure to sections of an image of an curved surface of observation. A change in measured grain size from one region of the image to another may be due to a change in

(a)



(b)

Figure 5-8 **Three circle (Abrams) procedure applied to macro-etched section of turbine airfoil. Figure 5-8(a) marks the grain crossings detected using a three concentric circle test pattern. Figure 5-8(b) shows the analysis of the intensity profile corresponding to the middle circle of figure 5-8(a)**

actual grain size or to a change in the view angle. Determining actual grain size requires that the foreshortening effect of the view angle be explicitly accounted for at each image point.

In a planimetric procedure, which involves counting the number of grains in a known area of test specimen, one can explicitly account for the effect of foreshortening if one knows the view angle (i.e., the magnitude of the gradient) at each image point. The surface area corresponding to a test region R in the image is given by:

$$A = \iint_R \sec(e)\, dxdy$$

where, once again, it is assumed that the image is an orthographic projection scaled so that image coordinates (x,y) correspond to object coordinates (x,y)

In an intercept procedure, which involves counting the number of grains crossed by a one-dimensional curve on the test specimen, one can explicitly account for the effect of foreshortening if one knows both components of the gradient at each image point. The path length ds corresponding to a movement [dx,dy] in the image is given by:

$$ds^2 = [dx\ dy] \begin{bmatrix} p^2+1 & pq \\ pq & q^2+1 \end{bmatrix} \begin{bmatrix} dx \\ dy \end{bmatrix} \qquad (5.4.1)$$

Equation (5.4.1) has the interpretation that the component of ds in the direction of steepest descent is foreshortened by the factor cos(e) while the component of ds in the direction of constant range is unchanged (see Appendix A.5).

Thus, in order to interpret the measurements derived by applying the three circle (Abrams) procedure to sections of an image of an arbitrary curved surface of observation, it is necessary to have the viewer-centered

representation for surface shape developed in Chapter 2. Conversely, with such a representation, the interpretation becomes straightforward.

The macro-etched turbine vane used in the example above was prepared using a contrast etch. A contrast etch makes the surface reflectance of the casting nonhomogeneous. Thus, in order to determine the viewer-centered representation required by (5.4.1), it would be necessary to first view an unetched part or have the representation supplied externally. On the other hand, for parts prepared using a flat etch, these two operations can be combined if the method for determining the viewer-centered representation is modified to ignore the dark markings corresponding to the grain boundries.

Fortunately, for airfoils, additional simplifications can be exploited. Airfoils can be approximated as singly curved surfaces. There is negligible curvature in directions parallel to the leading and trailing edges. Directions orthogonal to the leading and trailing edges are either strictly convex, on the "upper" surface, or strictly concave, on the "lower" surface.

In this chapter, emphasis has been placed on a method for quantitatively estimating the average grain size of metals. In quality control, the requirements are often more qualitative. It should be clear, however, that this chapter has developed a foundation for implementing a variety of qualitative techniques. The key is to be able to relate measurements made in an image to the topography of the surface being viewed. Columnar grain can be detected by comparing grain size estimates using two sets of linear intercept procedures. One set would sample lines parallel to the leading and trailing edges. The other set would sample lines orthogonal to the leading and trailing edges. Any significant

discrepancy would indicate the presence of columnar grain.

A more local version of the three circle (Abrams) procedure can be used to asign a grain size estimate to each point in an image. Demarcation lines between regions of different grain size would then appear as boundary lines in this "image" of local grain size values.

FOOTNOTE TO CHAPTER 5:

1. ASTM Designation: E112 - 74 is equivalent to American National Standard Z30.9 of the American National Standards Institute and has been approved by the Department of Defense to replace methods 311.1 and 312 of Federal Test Method Standard 151b and for listing in the DoD Index of Specifications and Standards.

6. EXPERIMENTS RELATING REFLECTANCE DATA AND OBJECT RELIEF

This chapter discusses the experimental determination of a reflectance map and its relation to object relief. The goal is to better understand how to acquire the data required to apply reflectance map techniques to practical situations. There is little literature on the measurement of the surface reflectance function $\phi(i,e,g)$ required for image analysis. A recent publication attempts to establish standards for nomenclature and measurement of reflectance [Nicodemus *et al* 77]. The results presented here, however, are meant to be illustrative, not conclusive.

6.1 SPECIFYING SURFACE REFLECTANCE

Two distinct physical processes are responsible for the reflection of radiant energy at boundary surfaces. The first of these is the mirror-like *specular reflection*. The second of these is the matte reflection which is believed to occur due to multiple reflections at the boundaries of small particles of which the surface is composed. This second process will be referred to as *diffuse reflection*. The *reflecting power* of a material is defined by:

$$R = \frac{I}{I_0}$$

where I_0 is the intensity of incident radiant energy and I is the intensity of energy reflected by the medium. There is confusion over terminology. The convention is sometimes adopted that the reflecting power associated with a specular process is termed *reflectivity*, while the reflecting power associated with a diffuse process is termed *reflectance* [Wendlandt & Hecht 66]. Elsewhere, the term reflectance is used to cover both cases [Morgan 53], [Jenkins & White 57]. Fortunately, in the

determination of the reflectance map $R(p,q)$, no fundamental distinction need be made between specular and diffuse reflection. In this work, the term reflectance is used to cover both cases.

A complete knowledge of the reflectance properties of a particular sample involves the determination of the spatial and spectral distribution of the reflected radiation with respect to both intensity and state of polarization. Reflectance measurements are an important tool in optics and analytic chemistry. The dependence of the reflecting power of optically smooth materials on wavelength, polarization and angle of incidence can be used to determine the fundamental optical constants of the material. Reflectance spectroscopy extends this analysis to materials with nonsmooth microstructure. Such studies are useful for the analysis of powders of known particle size and shape ground from samples that cannot be analyzed using traditional spectroscopic techniques, and for chemical constituent analysis of compound substances [Wendlandt & Hecht 66]. But, reflecting power measures only the fraction of the incident intensity reflected and not its spatial distribution.

Making use of reflectance measurements in image analysis requires an explicit account of the spatial distribution of the reflected energy as a function of the *spatial distribution* of the incident energy. It takes two parameters to specify a direction. But, if the reflecting material is isotropic, then the three photometric angles i, e and g, defined in figure 2-2, are sufficient to characterize the spatial relationship between *incident and reflected energy*. Let I_0 be the intensity of incident radiant energy as before. Consider a surface element of size ds. The surface element ds reflects energy into the hemisphere defined by $e < \pi/2$. Let I_1 be the intensity reflected by the surface element ds in the direction of

the viewer measured per unit solid angle per unit surface area perpendicular to the emitted ray. I_1 is called the *radiance* of the surface element ds and determines how bright it will appear when seen from view angle e. The surface reflectance function $\phi(i,e,g)$ is defined by:

$$\phi(i,e,g) = \frac{I_1}{I_0} \qquad (6.1.1)$$

Equation (6.1.1) does not distinguish between rotations of the surface element ds about the surface normal (refer to figure 2-2). Such rotations do not change i, e or g. This is the sense in which the surface material must be isotropic. (6.1.1) does not include the dependence of surface reflectance on the wavelength of the incident illumination. Thus, to use the corresponding reflectance map R(p,q) in image analysis, it is necessary to determine $\phi(i,e,g)$ under the same illumination conditions and sensor transfer characteristics that *will obtain when objects are to be viewed.*

One method for determining a reflectance map which avoids these problems is to measure the intensities recorded in an image of an object of known shape. The reflectance map R(p,q) so determined will hold for objects made from the same material and viewed with the same imaging device and under the *same object* surface, light source and viewer geometry.

Another method for determining a reflectance map is to use a photo-goniometer to fix g and explicitly measure the dependence of reflectance on the two photometric angles i and e. Such measurements were carried out using an improvised photo-goniometer and a sample gray iron casting as the specimen material.

6.2 MEASURING THE REFLECTANCE OF CAST GRAY IRON

First, in measuring reflectance, it is important to be clear on what the measurements are measurements of. Other definitions of surface reflectance arise which differ from the $\phi(i,e,g)$ defined in (6.1.1) by constant factors or by additional terms of cos(e) and/or cos(i). These distinctions can be made clear if one carefully analyzes the measurement situation.

Consider the experimental situation depicted in figure 6-1. Here, there is a distant source producing a narrow collimated beam of incident illumination. Suppose the incident intensity is I_0 per unit area measured perpendicular to the incident beam. Suppose the cross-sectional area of the incident beam is dA. In this situation, the total radiant flux incident on the surface is equal to $I_0 dA$, which is independent of i. Let the corresponding area of surface illuminated by the incident beam be ds. Then, dA corresponds to a foreshortened view of ds as seen by the source so that $ds = \cos(i)^{-1} dA$. Suppose that the receptive field of the distant detector is large compared to ds so that all of the flux reflected in the direction of the viewer is captured by the detector. Let F be the flux measured by the detector and let I_1 denote its intensity measured per unit solid angle per unit area perpendicular to the emitted beam. Then, $F = I_1 \cos(e) ds$ so that:

$$\phi(i,e,g) = \frac{I_1}{I_0} = \frac{F}{I_0 \cos(e) ds} = \frac{F \cos(i)}{I_0 \cos(e) dA} \qquad (6.2.1)$$

(6.2.1) relates measurements obtained in the experimental situation depicted in figure 6-1 to the definition of the surface reflectance function given in (6.1.1). I_0 and dA are constant. Multiplying the detector measurement F by the corresponding cos(i)/cos(e) produces a

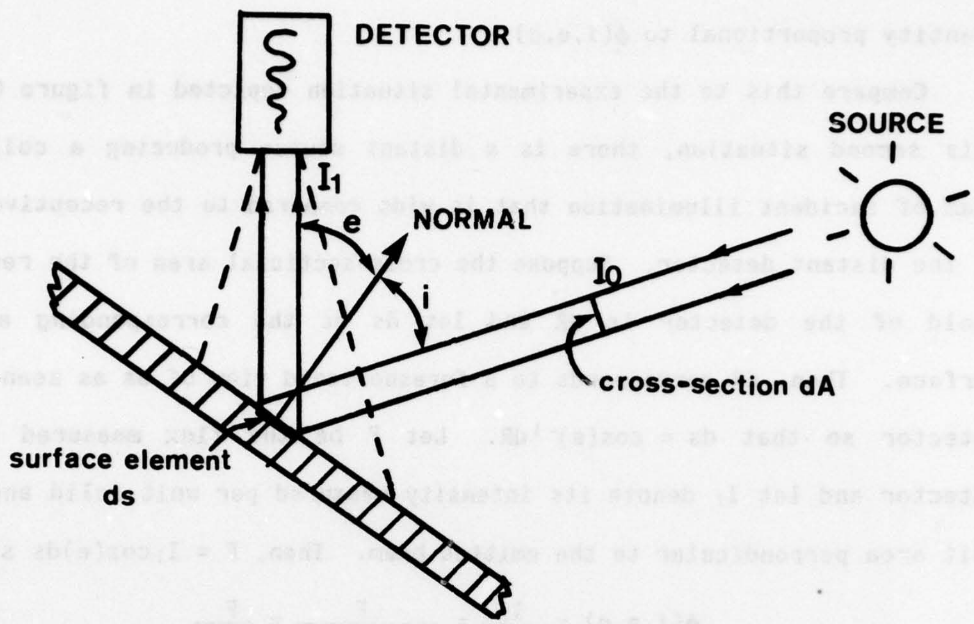


Figure 6-1 Measurement situation corresponding to a collimated beam of incident illumination which is narrow compared to the receptive field of the detector.

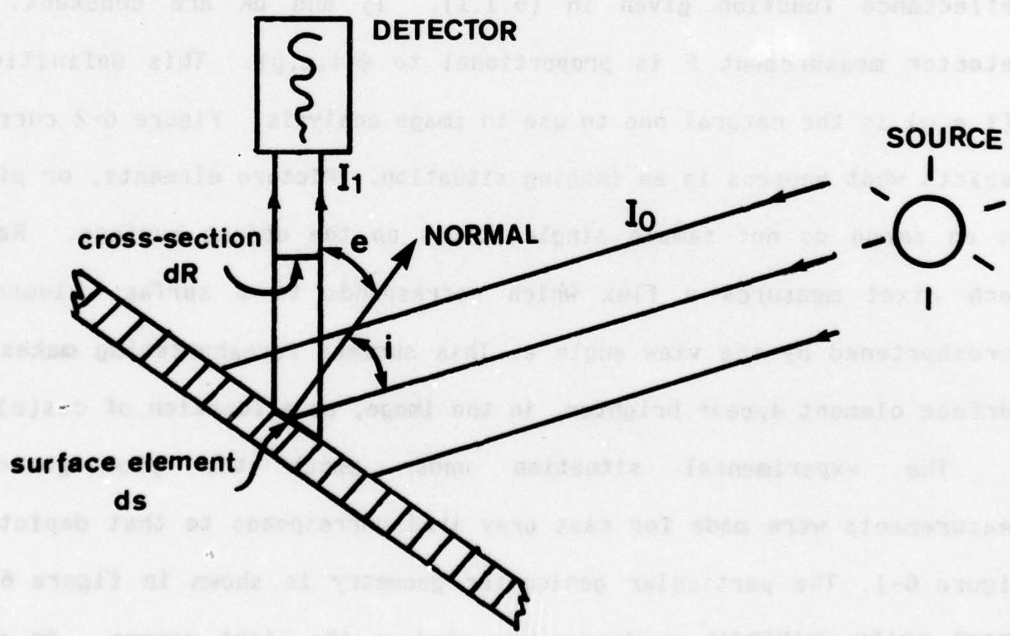


Figure 6-2 Measurement situation corresponding to a collimated beam of incident illumination which is wide compared to the receptive field of the detector.

quantity proportional to $\phi(i,e,g)$.

Compare this to the experimental situation depicted in figure 6-2. In this second situation, there is a distant source producing a collimated beam of incident illumination that is wide compared to the receptive field of the distant detector. Suppose the cross-sectional area of the receptive field of the detector is dR and let ds be the corresponding area of surface. Then, dR corresponds to a foreshortened view of ds as seen by the detector so that $ds = \cos(e)^{-1} dR$. Let F be the flux measured by the detector and let I_l denote its intensity measured per unit solid angle per unit area perpendicular to the emitted beam. Then, $F = I_l \cos(e) ds$ so that:

$$\phi(i,e,g) = \frac{I_l}{I_0} = \frac{F}{I_0 \cos(e) ds} = \frac{F}{I_0 dR} \qquad (6.2.2)$$

(6.2.2) relates measurements obtained in the experimental situation depicted in figure 6-2 directly to the definition of the surface reflectance function given in (6.1.1). I_0 and dR are constant. The detector measurement F is proportional to $\phi(i,e,g)$. This definition of $\phi(i,e,g)$ is the natural one to use in image analysis. Figure 6-2 correctly depicts what happens in an imaging situation. Picture elements, or *pixels*, in an image do not sample single points on the object surface. Rather, each pixel measures a flux which corresponds to a surface element ds foreshortened by the view angle e. This surface foreshortening makes each surface element appear brighter, in the image, as a function of cos(e).

The experimental situation under which the photo-goniometer measurements were made for cast gray iron corresponds to that depicted in figure 6-1. The particular goniometer geometry is shown in figure 6-3. A LEITZ PRADO UNIVERSAL projector was used as the light source. An opaque slide containing a small pinhole was projected to produce a collimated beam

of about 1 cm in diameter. A UNITED DETECTOR TECHNOLOGY, INC. 80X OPTO-METER fitted with a tele-photometer was used as the detector. The instrument was set to measure flux (in watts). At a distance of about 2 m, the receptive field of the tele-photometer subtended about 8 cm of sample surface. A small planar section of specimen material, about 3 cm by 8 cm, was mounted at the origin of the XYZ coordinate system. The phase angle g is the angle between the light source and the tele-photometer. For a particular set of measurements, the phase angle g was fixed. The view angle e and the angle of incidence i were varied by rotating the sample about the x-axis and y-axis.

To generate a reflectance map, gradient coordinates p and q must be defined in terms of the goniometer geometry shown in figure 6-3. Two anomalies of the goniometer geometry must be dealt with. First, the positions of the source and viewer have been interchanged with respect to the standard imaging geometry used to define the reflectance map in Chapter 2.3. Vector [0,0,-1] points at the source while vector [sin(g),0,-cos(g)] points at the viewer. Second, with the particular mechanical arrangement used, the x-axis and y-axis are coupled. That is, rotations about each axis are not independent. Let θ_x be the initial angle of rotation about the x-axis. Let θ_y be the subsequent angle of rotation about the y-axis. Specifying the order of rotation is necessary. For the results presented here, θ_x is set first followed by θ_y. Measurements were made in increments of 5° in both θ_x and θ_y

If one first rotates θ_x about the x-axis and then rotates θ_y about the y-axis, the equivalent gradient coordinates describing the resultant surface orientation are given by:

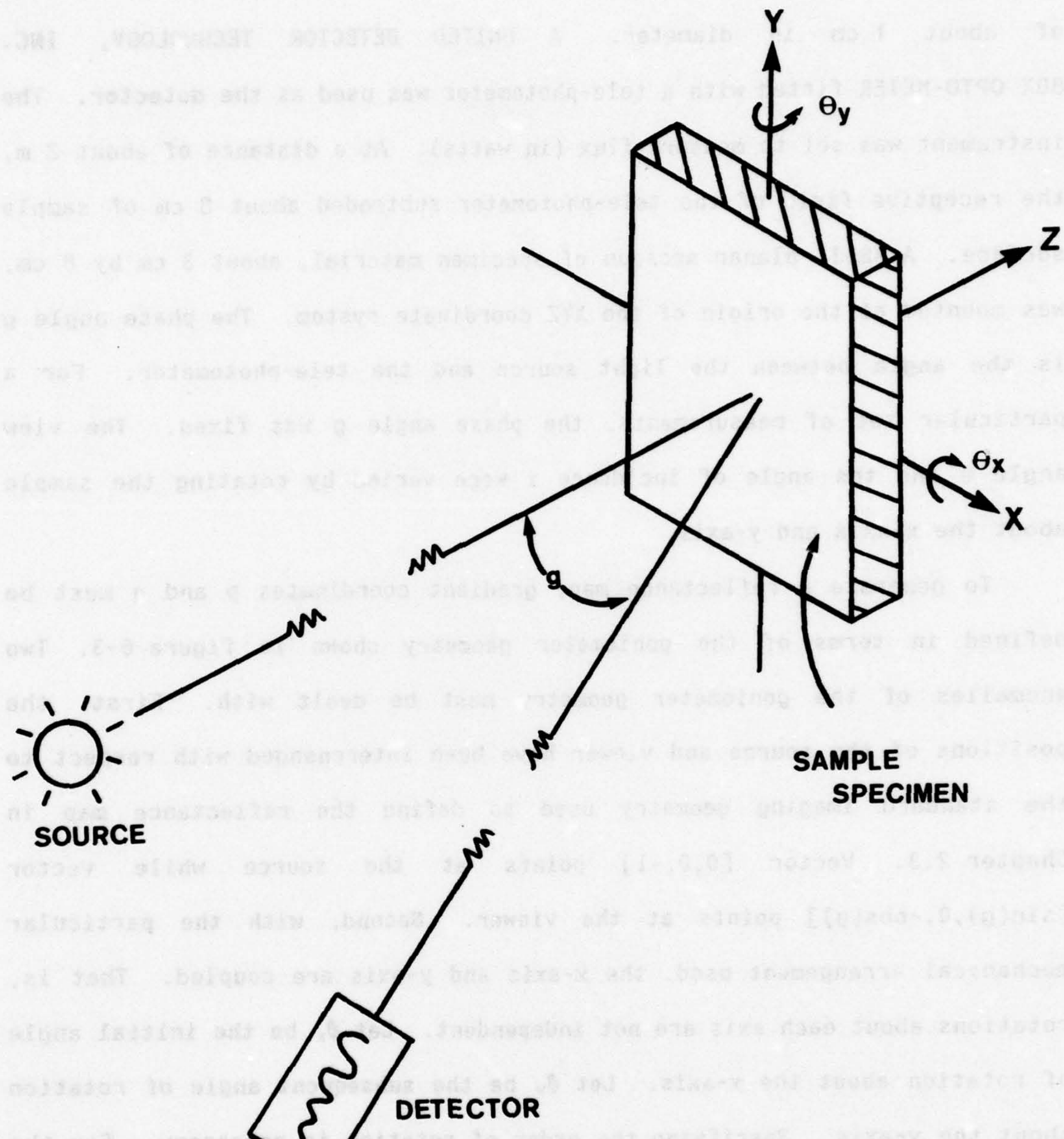


Figure 6-3 Photo-goniometer geometry.

$$p = \tan(\theta_y - g)$$

$$q = \frac{\tan(\theta_x)}{\cos(\theta_y - g)}$$

In this scheme, the position of the light source is given by:

$$p_s = -\tan(g)$$

$$q_s = 0$$

For the measurements obtained, the phase angle g was fixed at $g = 35^o$. To determine the reflectance map required for image analysis, the photo-goniometer measurements were multiplied by the "correction factor" cos(i)/cos(e), as indicated in (6.2.1). Subsequently, the measurements were normalized so that the maximum reflectance value was 1.0. The results are presented as a contour plot. Figure 6-4 plots these contours, spaced 0.1 units apart. The box enclosing the contours delimits that portion of gradient space which was measured. This limit reflects constraints imposed by the size of the specimen and the mechanical arrangement of the photo-goniometer.

Reflection from the surface of a metal is highly specular. Specular reflection from an optically smooth surface is easy to characterize. For such reflection, $\phi(i,e,g)$ is zero except when e = i and i, e and g lie in the same plane. Then, $\phi(i,e,g)$ is one. The viewing direction for which e = i and i, e and g lie in the same plane is called the *perfect specular direction*. As surface roughness increases, highly specular materials will reflect light in directions away from the perfect specular direction. For any viewing direction e, one can determine the angle between e and the perfect specular direction. This angle is called the *off-specularity angle s*. It can be shown that

$$\cos(s) = 2\cos(i)\cos(e) - \cos(g) \qquad (6.2.3)$$

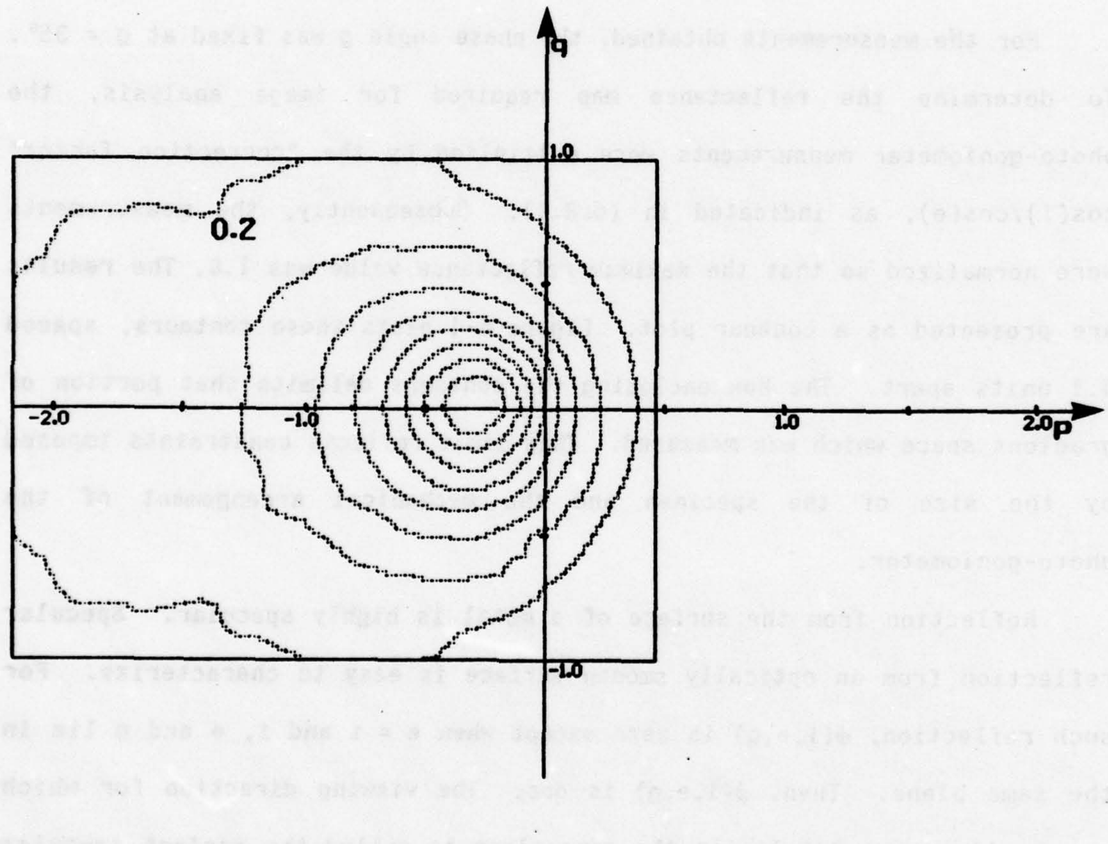


Figure 6-4 Measured reflectance map for cast gray iron with light source at p_s = -0.7 and q_s = 0.0 (contours are spaced 0.1 units apart). The solid box marks the region of gradient space over which measurements were made.

Figure 6-5 plots contours of constant cos(s) for $p_s = -\tan(35^\circ)$ $q_s = 0$, spaced 0.1 units apart. Contours of constant cos(s) are circles in gradient space. In general, if the light source is positioned at (p_s, q_s), then the contour $\cos(s) = k \geq 0$ is a circle in gradient space centered at (cp_s, cq_s) and of radius r where:

$$c = \frac{\cos(g)}{k + \cos(g)}$$

and

$$r^2 = \frac{1 - k^2}{[k + \cos(g)]^2}$$

For most surfaces, measurements suggest that both specular and diffuse reflection is always present, the relative proportion of each depending on the nature of the material [Wendlandt & Hecht 66]. By combining a term in cos(s), to account for the specular component, and a term in cos(i), to account for the diffuse component, it is possible to model different surface materials.

A good approximation for many materials is achieved by letting the surface reflectance function be:

$$\phi(i,e,g) = t\frac{(n+1)}{2}\cos(s)^n + (1-t)\cos(i) \qquad (6.2.4)$$

where t lies between 0 and 1 and determines the fraction of incident light reflected specularly, a parameter which models the optical properties of the material, and n determines the sharpness of the specular peak, a parameter which models the surface microstructure [Horn 77]. Figure 6-6 plots the reflectance map obtained using (6.2.4) with $p_s = -0.7$, $q_s = 0$, $n = 3$ and $t = 0.6$. There is good qualitative agreement between figure 6-6 and the data shown in figure 6-4. This agreement, however, has been verified for only one phase angle and over a limited region of gradient space.

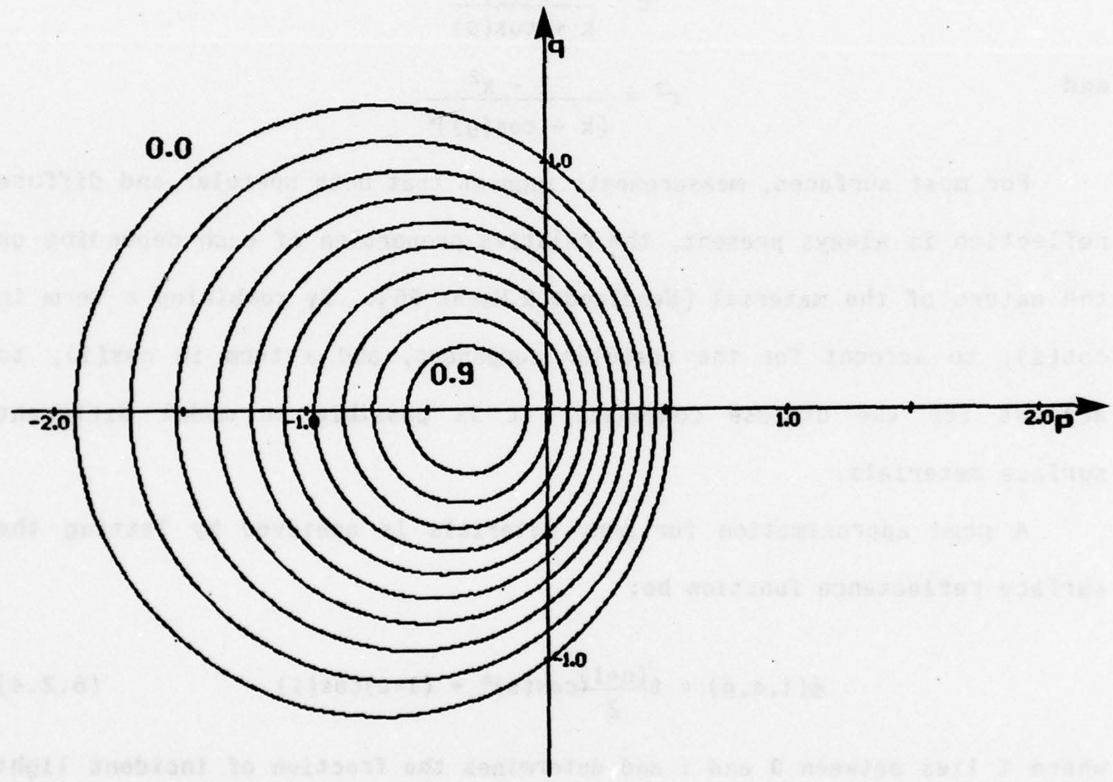


Figure 6-5 Contours of constant cos(s) for light source at p_s = -0.7 and q_s = 0.0 (contours are spaced 0.1 units apart).

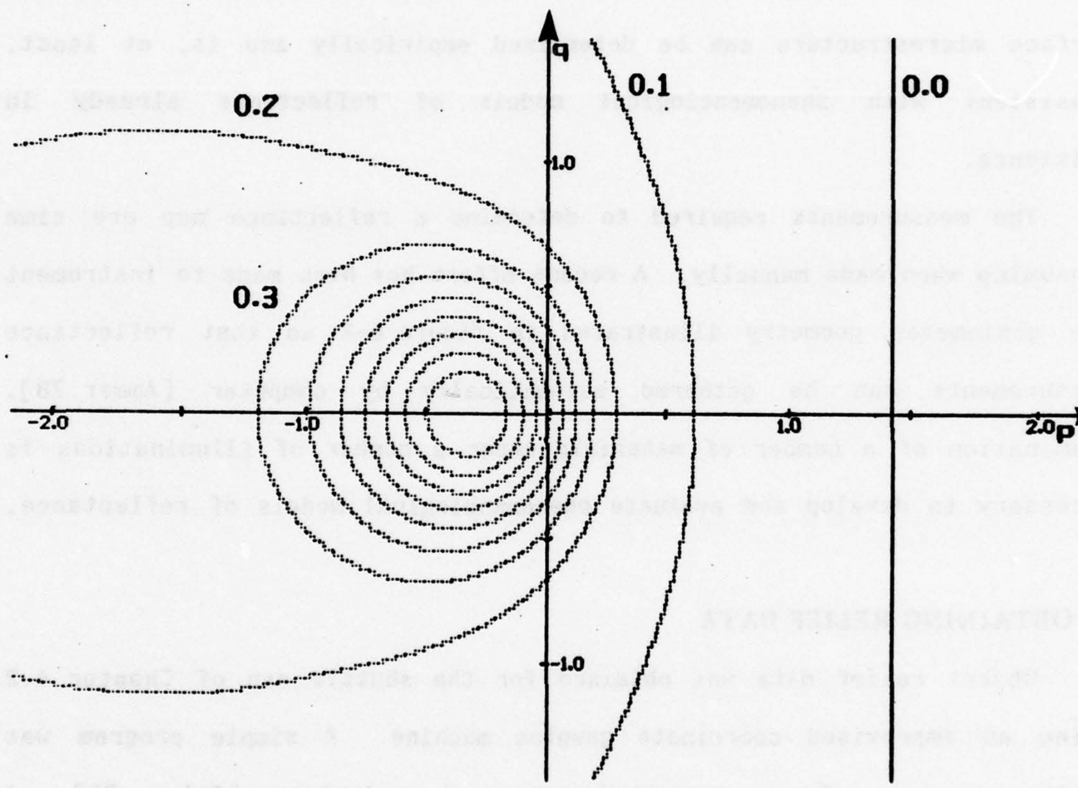


Figure 6-6 Phenomenological model of reflectance map for gray cast iron with light source at p_s = -0.7 and q_s = 0.0 (contours are spaced 0.1 units apart).

It is important to point out that (6.2.4) is a phenomenological model and is not based on any optical theory or actual reflectance measurements. The purpose is not to produce closed form expressions for the surface reflectance of gray iron. At best, such an expression would vary with alloy composition and sand grain size. Rather, the purpose has been to demonstrate that a reflectance map for a particular alloy with a particular surface microstructure can be determined empirically and is, at least, consistent with phenomenological models of reflectance already in existence.

The measurements required to determine a reflectance map are time consuming when made manually. A recent effort has been made to instrument the goniometer geometry illustrated in figure 6-3 so that reflectance measurements can be gathered automatically by computer [Ammar 78]. Examination of a number of materials under a number of illuminations is necessary to develop and evaluate phenomenological models of reflectance.

6.3 OBTAINING RELIEF DATA

Object relief data was obtained for the shuttle eye of Chapter 4.2 using an improvised coordinate gauging machine. A simple program was written to use a force controlled mechanical manipulator [Silver 73]. A needle probe was inserted between the manipulator grippers and the surface scanned at a resolution of 0.01 inches in x and y. At each (x,y) point, the manipulator was programmed to descend until it "touched" the surface. The height of first touch was recorded to create a range file. Figure 6-7 plots, from four different viewing directions, the range data obtained for the shuttle eye. Figure 6-8 is an image of this shuttle eye synthesized from the range data and the reflectance map determined above.

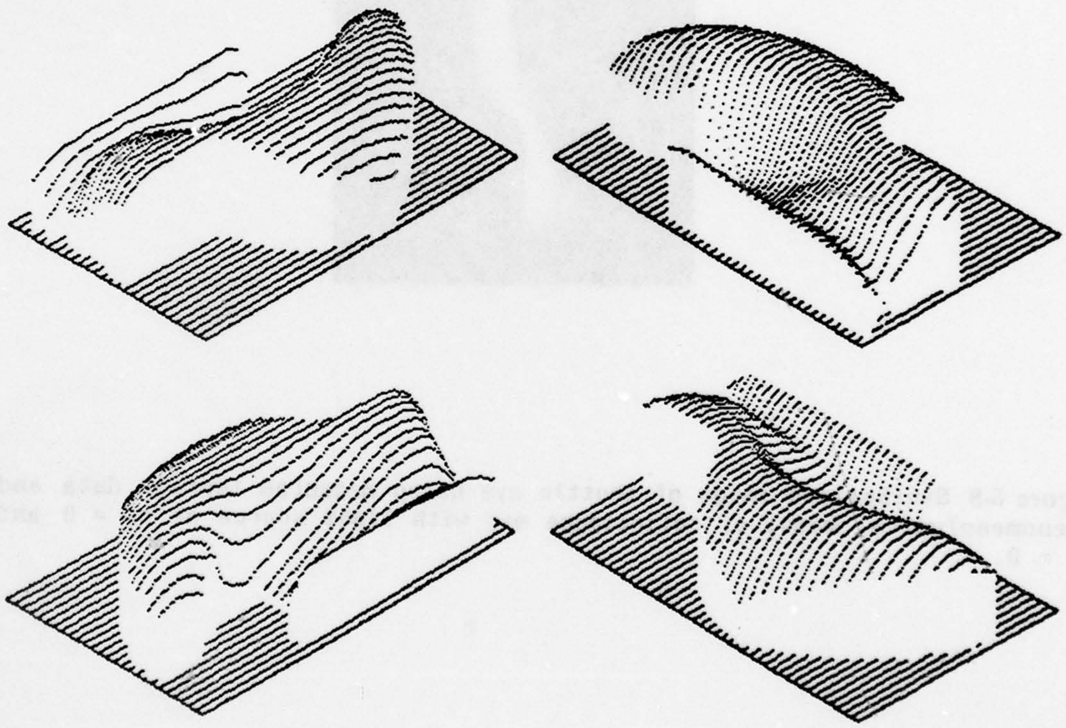


Figure 6-7 Plot of measured terrain data for shuttle eye (from four views).

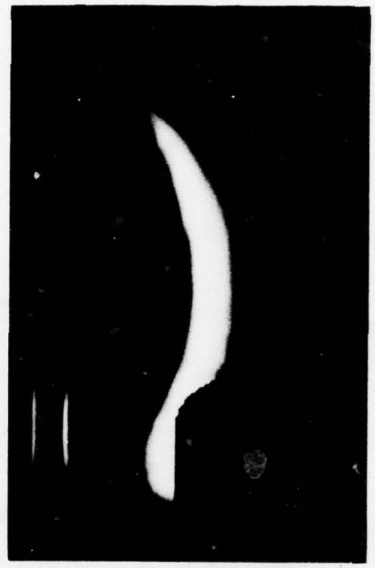


Figure 6-8 Synthesized image of shuttle eye using measured terrain data and phenomenological model of reflectance map with light source at $p_s = 0$ and $q_s = 0$.

7. CONCLUDING REMARKS

This chapter summarizes the research in terms of the goals initially set forth and the results achieved. The work had two goals: one theoretical and one practical. The theoretical goal was to understand how to interpret image intensity in terms of the underlying surface topography of the objects being imaged. The practical goal was to explore how machine vision systems could be applied to the problem of inspecting surface defects in metal castings.

With these goals in mind, the research naturally split into two parts. The first part consisted of a basic investigation into the problem of determining object relief directly from image intensity. Image analysis is a hard problem because many factors influence the intensity values recorded in an image. Image analysis requires that properties of image intensity be related to properties of the objects being imaged. This work has explored the fundamental nature of this computation. [Horn 75] [Horn 77] provided a framework for modeling both the geometry and radiometry of the image-forming process. A key observation is that, in a mathematical sense, the problem of computing object relief locally at each image point is underdetermined. In order to assign a global interpretation to an image, it necessary to invoke additional assumptions about the nature of the object surface in view.

The present work has extended the pioneer work of Horn by providing a mechanism for representing assumptions about surface curvature and by demonstrating how this representation can be exploited in methods for computing object relief from image intensity. The image Hessian has been defined and used to represent surface curvature. Assumptions about surface curvature map naturally into properties of the image Hessian.

In its theoretical contribution, the research is evolutionary rather than revolutionary. Dealing with arbitrary surfaces remains a difficult problem. In general, even if the reflectance map R(p,q) is known explicitly, there is no unique surface z = f(x,y) corresponding to a given image I(x,y). Yet, people provide consistent interpretations of images. One unexplored area of research is to try to formulate constraints on the computation which lead to the solutions people most often assign. An idea, for example, is to compute the surface z = f(x,y) which requires the least surface area to account for the observed change in intensity. Such constraints, while interesting, fall outside the scope of the present work. In determining what can be computed directly from image intensity, it is also useful to point out what cannot.

At the same time, many examples have been presented in which image analysis simplifies. Surfaces which are singly curved and surfaces which are solids of revolution (i.e., have a circular cross-section) are particularly easy to deal with. Such simplifications arise naturally in practical situations since there are few design techniques and few fabrication methods for manufacturing parts with arbitrary doubly curved surfaces. The design of curved surfaces has always presented difficulties to the engineer. Some types of curved shapes, such as cylinders, spheres, surfaces of revolution and singly curved surfaces can be represented on a two-dimensional drawing with considerable precision. Wherever possible the engineer constrains curved portions of the design to one of these types. There are few drafting techniques for representing curves in a drawing which are not planar curves in space. Not surprisingly, such simplifications also seem to arise naturally in the interpretation humans assign to drawings [Marr 77a].

The second part consisted of a practical demonstration of a working inspection system. Having discussed the needs of casting inspection with various representatives of the castings industry, a particular inspection problem was chosen. The industry standard for estimating the average grain size of metals was implemented and demonstrated on a sample turbine vane. The image analysis portion of the program incorporated existing ideas in machine vision [Marr 76], [Lozano-Perez 77]. For a standard part surface and a standard method of part presentation, analysis of intensity profiles is sufficient to verify grain structure for equiaxed turbine blades and vanes. This demonstrates that machine vision systems can tackle existing inspection problems. But, flexible inspection systems must be able to interpret measurements made from an image in terms of the underlying surface topography of the part being inspected.

This example relates to an important fact of industrial automation. Current large volume automation incorporates a wide variety of special purpose "tricks" to accomplish specific tasks of parts feeding, assembly and inspection. Unfortunately, there is little or no general theory of how to do this. This presents no problem when the volume of production justifies an investment in special purpose trickery. But, it also means that low to mid volume production can not profitably be automated.

Machine vision is an attractive interface to existing production facilities. In the immediate future, industrial applications of machine vision can be expected to take the form of special purpose tricks. This is reasonable if the domain of interest is constrained so that special purpose tricks will work. At the same time, one would like to have an underlying theory of image formation. At the very least, such a theory would help in the design of special purpose tricks.

Analyzing surface defects in green sand mold castings seems to require more than special purpose tricks. One idea might be to compare the image of each casting against the image of a casting known to be free of defects. Parts whose intensities match to within a specified tolerance would be accepted while those outside the specified tolerance would be rejected. This idea is like a method of inspection used in many foundries. Castings can be weighed cheaply. If the casting does not weigh as much as it should, then there must be shrinkage voids or regions of high porosity (i.e., holes) inside the casting. This is a convenient way to find obvious internal defects without the high cost of a more quantitative analysis (i.e., radiographic or ultrasonic inspection). This method is successful since there is a simple linear relationship between measured weight and the mass of metal missing from the part.

Unfortunately, in comparing images, there is no such simple relationship between differences in intensity and surface defects. A particular surface feature maps into many possible features of intensity depending on the position of the light source and and on the surface-viewer geometry. Simple difference detectors will fail unless the measurements made can be interpreted in the context of the part as a whole.

One way to relate image intensity to part geometry is to match the image to a known model of the part. Two-dimensional models are useful for determining part identity, position and orientation from boundary information [Perkins 77]. Interpreting local features of intensity as defects on sections of smooth surface, however, requires a three-dimensional surface model. The surface function $z = f(x,y)$ is such a three-dimensional model.

For a surface with homogeneous optical properties, photometric stereo is a straightforward method for determining the underlying surface $z = f(x,y)$. This technique seems particularly suited to casting inspection. The two properties which determine the surface photometry of a casting, namely alloy composition and surface microstructure, are carefully controlled. The production engineer is also free to standardize the illumination and viewing geometry at each inspection station.

Typical surface defects in castings (eg. pinholes, cold shuts, cracks and hot tears) manifest themselves as properties of surface topography that could not have been the result of an intended casting operation. To do the kind of first visual inspection required in a typical batch-oriented foundry, it is probably not necessary to have a precise model of each part geometry. Imagine an inspection system for finding cold shut defects in green sand mold castings. Photometric stereo could be used to determine the shape of the casting. Subsequent analysis would look for variations in intensity at concave boundaries between large sections of the casting that indicate a lapping or layering of the surface.

This research has explored ways of determining the surface function $z = f(x,y)$ directly from image intensity. Exciting new techniques in image analysis are possible once the image $I(x,y)$ has been aligned with the surface function $z = f(x,y)$. Once alignment has been achieved, it becomes possible to distinguish effects due to illumination and varying surface slope from effects due to varying surface cover [Nitzan *et al* 77] [Barrow & Tenenbaum 78] [Horn & Bachman 77]

REFERENCES

Agin, G. J. and Binford, T. O. (1973), "Computer Description of Curved Objects", in *Proceedings of the 3rd International Joint Conference on Artificial Intelligence*, pp. 629-640, August 1973.

Ammar, M. H. (1978), "A Real-Time System For Determining Reflectance Maps", S.B. Thesis, Dept. of Elect. Eng. & Comp. Sci., M.I.T., May 1978.

ASM (1970), *Metals Handbook (8th Edition), Vol. 5, Forging and Casting*, American Society for Metals, Metals Park, Ohio, 1970.

ASTM (1976), ASTM Designation: E 112-74, "Standard Methods for Estimating the Average Grain Size of Metals", in *1976 Annual Book of ASTM Standards*, Part 11, pp. 208-240, American Society for Testing and Materials, 1976.

Barrow, H. G. & Tenenbaum, J. M. (1978), "Recovering Intrinsic Scene Characteristics from Images", in *Computer Vision Systems*, A. Hanson and E. Riseman (ed.), Academic Press, N.Y., 1978.

Beck, J. (1974), *Surface Color Perception*, Cornell University Press, 1974.

Binford, T. O. and Horn, B. K. P. (1973), "The Binford-Horn Line Finder", AI MEMO 285, AI Laboratory, M.I.T., December 1973.

Caulfield, H. J. (1976), "Three Dimensional Range Finding", *International Optical Computing Conference*, Digest of Papers, pp. 78-79, IEEE Catalog No. 76CH1100-7C, August 31-September 2, 1976.

Clowes, M. B. (1971), "On Seeing Things", in *Artificial Intelligence*, Vol 2, pp. 79-112, 1971.

Draper (1973), "Annual Progress Report No. 2 for the Development of Multi-Moded Remote Manipulator Systems (Period Ending January 1973)", C. S. Draper Laboratory, M.I.T., March 1973.

Ejiri, M., Uno, T., Mese, M., and Ikeda, S. (1973), "A Process for Detecting Defects in Complicated Patterns", *Computer Graphics and Image Processing*, Vol 2, No. 3/4, pp. 326-329, December 1973.

Falk, G. (1972), "Interpretation of Imperfect Line Data as a Three-dimensional Scene", in *Artificial Intelligence*, Vol 3, pp. 101-144, 1972.

Freuder, E. C. (1976), "A Computer System for Visual Recognition Using Active Knowledge", AI TR 345, AI Laboratory, M.I.T., June 1976.

Guzman, A. (1968), "Computer Recognition of Three-Dimensional Objects in a Visual Scene", MAC TR-59 (Thesis), Project MAC, M.I.T., December 1968.

Horn, B. K. P. (1968), "Focussing", AI-MEMO 160, AI Laboratory, M.I.T., May 1968.

Horn, B. K. P. (1975), "Obtaining Shape from Shading Information", in *The Psychology of Computer Vision*, P. H. Winston (ed.), McGraw-Hill, pp. 115-155, 1975.

Horn, B. K. P. (1977), "Understanding Image Intensities", in *Artificial Intelligence*, Vol 8, pp. 201-231, 1977.

Horn, B. K. P. & Bachman, B. L. (1977), "Using Synthetic Images to Register Real Images with Surface Models", AI Memo 437, M.I.T. AI Laboratory, August 1977.

Huffman, D. A. (1971), "Impossible Objects as Nonsense Sentences", in *Machine Intelligence 6*, R. Meltzer and D. Michie (ed.), Edinburgh University Press, pp. 295-323, 1971.

Huffman, D. A. (1975), "Curvature and Creases: A Primer on Paper", *Proceedings of Conference on Computer Graphics, Pattern Recognition and Data Structures*, pp. 360-370, May 1975.

Jenkins, F. A. and White, H. E. (1957), *Fundamentals of Optics*, (3rd edition), McGraw-Hill, 1957.

Johnston, A. R. (1973), "Infrared Laser Rangefinder", Report NPO-13460, Jet Propulsion Laboratory, Pasadena, California, August 1973.

Kepr, B. (1969), "Differential Geometry", in *Survey of Applicable Mathematics*, K. Rektorys (ed.), The M.I.T. Press, pp. 298-373, 1969.

Lozano-Perez, T. (1977), "Parsing Intensity Profiles", *Computer Graphics and Image Processing*, Vol 6, No. 1, February 1977.

Luenberger, D. G. (1973), *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, 1973.

Mackworth, A. K. (1973), "Interpreting Pictures of Polyhedral Scenes", in *Artificial Intelligence*, Vol 4, pp. 121-137, 1973.

Mangasarian, O. L. (1969), *Nonlinear Programming*, McGraw-Hill, 1969.

Marr, D. (1976), "Early Processing of Visual Information", *Phil. Trans. Roy. Soc. B.*, 275, pp. 483-524, 1976.

Marr, D. (1977a) "Analysis of Occluding Contour", *Proc. R. Soc. Lond. B.*, 197 pp. 441-475, 1977.

Marr, D. (1977b) "Representing Visual Information", AI-MEMO 415, AI Laboratory, M.I.T., May 1977.

Marr, D. and Poggio, T. (1976), "Cooperative Computation of Stereo Disparity", in *Science 194*, pp. 283-287.

Moore, J. T. (1968), *Elements of Linear Algebra and Matrix Theory*, McGraw-Hill, 1968.

Morgan, J. (1953), *Introduction to Geometrical and Physical Optics*, McGraw-Hill, 1953.

Nevatia, R. & Binford, T. O. (1973), "Structured Descriptions of Complex Objects", in *Proceedings of the 3rd International Joint Conference on Artificial Intelligence*, pp. 641-645, August 1973.

Nicodemus, F. E., Richmond, J. C. & Hisa, J. J. (1977), "Geometrical Considerations and Nomenclature for Reflectance", NBS Monograph 160, National Bureau of Standards, Washington, D.C., October 1977.

Nitzan, D., Brain, A. E. & Duda, R. O. (1977), "The Measurement and Use of Registered Reflectance and Range Data in Scene Analysis", *Proc. IEEE*, pp. 206-220, Feb. 1977.

Perkins, W. A. (1977), "Model-Based Vision System for Scenes Containing Multiple Parts", in *Proceedings of IJCAI-77*, pp. 678-684, August 1977.

Phong, B. T. (1975), "Illumination for Computer Generated Pictures", *CACM, 18* pp. 311-317, 1975.

Plass, G. N., Kattawar, G. W. and Guinn Jr., J. A. (1977), "Isophotes of Sunlight Glitter on a Wind-Ruffled Sea", *Applied Optics*, pp. 643-653, March 1977.

Rindfleisch, T. (1966), "Photometric Method for Lunar Topography", *Photogrammetric Engineering*, pp. 262-276, March 1966.

Roberts, L. G. (1965), "Machine Perception of Three-Dimensional Solids", in *Optical and Electro-optical Information Processing*, J. T. Tippett *et al* (ed.), M.I.T. Press, pp. 159-197, 1965.

Rosenfeld, A., Hummel, R. A. and Zucker. S. W. (1976), "Scene Labelling by Relaxation Operations", *IEEE Transactions on Systems, Man and Cybernetics, SMC-6*, pp. 420-433.

Shirai, Y. (1975), "Analyzing Intensity Arrays Using Knowledge About Scenes", in *The Psychology of Computer Vision*, P. H. Winston (ed.), McGraw-Hill, pp. 93-113, 1975.

Silver, D. (1973), "The Little Robot System", AI MEMO 273, AI Laboratory, M.I.T., January 1973.

Sullivan, C. P. (1976), "Unidirectionally Solidified Superalloy Airfoils", *Foundry M&T*, pp. 118-121, Jan. 1976.

van Diggelen, J. (1951), "A Photometric Investigation of the Slopes and the Heights of the Ranges of Hills in the Maria of the Moon", *Bulletin of the Astronomical Institutes of the Netherlands*, Vol. XI, No. 423, July 1951.

Wendlandt, W. W. and Hecht, H. G. (1966), *Reflectance Spectroscopy*, Interscience Publishers, John Wiley & Sons, 1966.

Waltz, D. (1975), "Understanding Line Drawings of Scenes with Shadows", in *The Psychology of Computer Vision*, P. H. Winston (ed.), McGraw-Hill, pp. 19-91, 1975.

Will, P. M. and Pennington, K. S. (1971), "Grid Coding: A Preprocessing Technique for Robot and Machine Vision", *Advance Papers of the 2nd International Joint Conference on Artificial Intelligence*, pp 66-70, Spetember 1971.

Winston, P. H. (1973), "The M.I.T. Robot", in *Machine Intelligence 7*, B. Meltzer and D. Michie (ed.), Edinburgh University Press, pp. 431-463, 1973.

Winston, P. H. (1975), "Learning Structural Descriptions From Examples", in *The Psychology of Computer Vision*, P. H. Winston (ed.), McGraw-Hill, pp. 157-209, 1975.

Young, I. T. (1969), "Automated Leukocyte Recognition", PhD Thesis, M.I.T., 1969.

APPENDIX A: MATHEMATICAL DETAILS

The purpose of this appendix is to explore more formally the question of how physical constraints on the object surface $z = f(x,y)$ constrain possible solutions to the basic image-forming equation $I(x,y) = R(p,q)$. Intuitive reasoning about smooth surfaces and gradient space needs to be augmented by some formal definitions and theorems. The mathematics presented in this appendix is not new. The pertinent definitions and theorems arise from work in linear algebra, convex analysis, differential geometry and nonlinear programming. The goal here is to apply this mathematics to the problem of interpreting image intensities. The required formal results are merely summarized. Technical details of proofs have been omitted. For those details, the reader is referred to [*Mangasarian* 69], [*Moore* 68], [Luenberger 73] and [Kepr 69].

In order to develop the required theorems, the notion of surface smoothness must be made precise.

> Definition. Let $f(x)$ be a real-valued function defined on an open set $\Gamma \subset \mathbb{R}^n$. Then the (hyper)surface described by the equation $z = f(x)$ is *smooth over* Γ if $f(x)$ is twice differentiable with continuous second partial derivatives at every $x \in \Gamma$.

It is also important to deal explicitly with surface curvature. The Hessian matrix H is the generalization to $\mathbb{R}^n$ of the concept of the curvature of a function of a real variable.

> Definition. Let $f(x)$ be a real-valued function defined on an open set $\Gamma \subset \mathbb{R}^n$ and let $x^* \in \Gamma$. Then, the $n \times n$ matrix $H = \nabla^2 f(x^*)$, whose ij^{th} element is given by:
>
> $$[\nabla^2 f(x^*)]_{ij} = \frac{\partial^2 f(x^*)}{\partial x_i \partial x_j}$$
>
> is called the *Hessian (matrix) of* $f(x)$ *at* x^*.

The following theorem establishes that the Hessian matrix is symmetric for smooth surfaces.

> **Theorem A.1** Let $f(x)$ be a real-valued function defined on an open set $\Gamma \subset \mathbb{R}^n$ and let $x^* \in \Gamma$. If the (hyper)surface described by the equation $z = f(x)$ is smooth at x^*, then
>
> $$\frac{\partial}{\partial x_i}\frac{\partial f(x^*)}{\partial x_j} = \frac{\partial}{\partial x_j}\frac{\partial f(x^*)}{\partial x_i}$$
>
> *Corollary* A.1 If $z = f(x)$ is *smooth* at x^*, then the Hessian matrix H exists at x^* and is symmetric.

It can now be shown that the Hessian matrix H underlies the model of the image forming process. Once again, recall that the assumptions used are that the viewer is distant, that the image projection is orthographic taking object point (x,y,z) onto image point (x,y) and that each object point receives the same incident illumination. The basic image-forming *equation* $I(x,y) = R(p,q)$ is *one equation* in the two unknowns p and q. By taking partial derivatives of this equation with respect to X and Y, two equations are obtained:

$$I_x = p_x R_p + q_x R_q$$
$$I_y = p_y R_p + q_y R_q$$

(subscripts are used to denote partial differentiation). Theorem A.1 says that, for a smooth surface, $p_y = q_x$. Thus, two equations are obtained in the three unknowns p_x, q_x and $p_y = q_x$ where:

$$p_x = \frac{\partial^2 f(x,y)}{\partial x^2} \qquad q_y = \frac{\partial^2 f(x,y)}{\partial y^2} \qquad p_y = q_x = \frac{\partial^2 f(x,y)}{\partial x \partial y}$$

These two equations can be written as the single matrix equation:

$$\begin{bmatrix} I_x \\ I_y \end{bmatrix} = \begin{bmatrix} p_x & q_x \\ p_y & q_y \end{bmatrix} \begin{bmatrix} R_p \\ R_q \end{bmatrix}$$

The relationship between the vector of first partial derivatives of the intensity function $[I_x, I_y]$ and the corresponding vector of first partial

derivatives of the reflectance map function $[R_p, R_q]$ is given by:

$$[I_x, I_y]^T = H\,[R_p, R_q]^T \qquad \text{(A.1)}$$

where $H = \nabla^2 f(x,y)$ is the Hessian matrix of second partial derivatives of the object surface $z = f(x,y)$.

First-order approximations, written in terms of differentials, relating a small movement $[dx, dy]$ in the image to the corresponding movement $[dp, dq]$ in gradient space are given by:

$$dp = p_x\,dx + p_y\,dy$$

$$dq = q_x\,dx + q_y\,dy$$

Again, these two equations can be written as the single matrix equation:

$$\begin{bmatrix} dp \\ dq \end{bmatrix} = \begin{bmatrix} p_x & p_y \\ q_x & q_y \end{bmatrix} \begin{bmatrix} dx \\ dy \end{bmatrix}$$

To a first approximation, the relationship between a small movement $[dx, dy]$ in the image and the corresponding movement $[dp, dq]$ in gradient space is also determined by the Hessian matrix H. It is given by the equation:

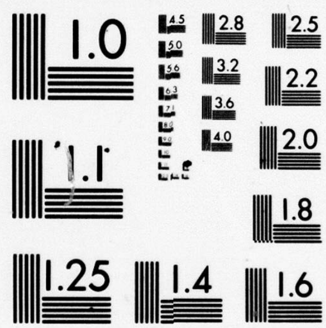
$$[dp, dq]^T = H\,[dx, dy]^T \qquad \text{(A.2)}$$

A few words of caution are in order. While equation (A.1) above is exact at any image point (x,y) and its corresponding gradient point (p,q), equation (A.2) is only approximate. The Hessian matrix H is a function of x and y. In the subsequent analysis, however, it is assumed that $[dx, dy]$ can be chosen small enough so that H can be considered constant over the interval (x,y) to $(x+dx, y+dy)$.

AD-A062 177 MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTE-ETC F/G 20/6
REFLECTANCE MAP TECHNIQUES FOR ANALYZING SURFACE DEFECTS--ETC.(U)
JUN 78 R. J. WOODHAM N00014-77-C-0389
UNCLASSIFIED AI-TR-457 N/L

3 OF 3
AD A
062177

END
DATE
FILMED
5-82
DTIC



MICROCOPY RESOLUTION TEST CHART

Suppose image point (x,y) is known to correspond to gradient point (p,q). If two linearly independent directions $[dx_1, dy_1]$ and $[dx_2, dy_2]$ and the corresponding $[dp_1, dq_1]$ and $[dp_2, dq_2]$ are known, then the Hessian matrix **H** is determined uniquely at (x,y). Indeed,

$$\mathbf{H} = \begin{bmatrix} dp_1 & dp_2 \\ dq_1 & dq_2 \end{bmatrix} \begin{bmatrix} dx_1 & dx_2 \\ dy_1 & dy_2 \end{bmatrix}^{-1}$$

A.1 HORN'S METHOD FOR OBTAINING SHAPE FROM SHADING INFORMATION

It is now possible to re-examine the method Horn developed for obtaining shape from shading information [Horn 75] [Horn 77]. The goal is to interpret Horn's method in terms of the image Hessian **H**.

The basic recipe for Horn's solution is as follows:

(i) Suppose image point (x,y) is known to correspond to a point (p,q) in gradient space. Then, the change in z = f(x,y) corresponding to a small movement [dx,dy] in the image is given by the first-order approximation:

$$dz = p\,dx + q\,dy$$

(ii) The new gradient point corresponding to the image point (x+dx,y+dy) is obtained by updating the current gradient (p,q) according to the equation:

$$[dp, dq]^T = \mathbf{H}\,[dx, dy]^T$$

Unfortunately, there is not enough information to determine the matrix **H**. The only constraint on **H** is that it satisfies the equation:

$$[I_x, I_y]^T = \mathbf{H}\,[R_p, R_q]^T$$

Note, however, that matrix multiplication is a linear operation. If [dx,dy] is chosen to be in the direction of $[R_p, R_q]$ then linearity is sufficient to guarantee that [dp,dq] will be in the direction of $[I_x, I_y]$. More precisely:

$$\text{If } [dx, dy] = [R_p, R_q]\,ds, \text{ then } [dp, dq] = [I_x, I_y]\,ds$$

Thus, by starting at a point (x,y) known to have gradient (p,q) and iterating the above two operations, a path in the image is traced out for which the corresponding gradients, and hence the corresponding relief profile on the object surface, can be determined. The catch is that an arbitrary direction for [dx,dy] cannot be chosen. [dx,dy] must be chosen in the direction $[R_p, R_q]$. The curves traced out on the surface in this fashion are called *characteristics* and their projections in the image plane are called *base characteristics*.

This result can also be interpreted geometrically. Choosing [dx,dy] to be in the direction $[R_p, R_q]$ means that a base characteristic is traced out that is always perpendicular to the contour of constant reflectance at the current (p,q). Similarly, the fact that the resulting [dp,dq] is in the direction $[I_x, I_y]$ means that the corresponding path traced out in gradient space is always perpendicular to the contour of constant intensity at the current (x,y). Figure A-1 illustrates this result.

The path the base characteristic traces out cannot be controlled. It depends on the particular object in view. Assumptions are required to link the characteristics traced out from different starting points.

A.2 THE IMAGE HESSIAN MATRIX

In this work, the requirement that [dx,dy] be chosen in a particular direction is relaxed. The reason is not to make the problem more difficult! Rather, it is to try to gain more control over the "path" that can be analyzed, independent of the object in view. To do this, the analysis of how properties of the object surface z = f(x,y) relate to properties of the image Hessian matrix H is extended. These properties of H constrain the set of [dp,dq]'s that can correspond to a particular

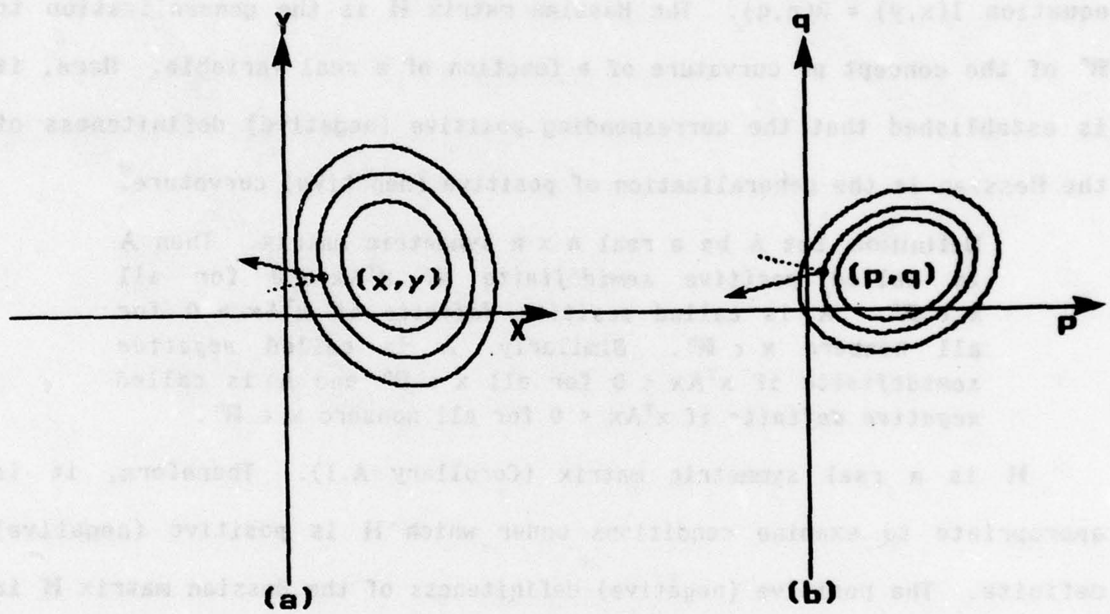


Figure A-1 Suppose image point (x,y) corresponds to gradient (p,q). In expanding a characteristic, movement in the image is in the direction normal to the contour of constant reflectance at (p,q). Similarly, movement in gradient space is in the direction normal to the contour of constant intensity at (x,y).

[dx,dy].

If H were known everywhere then the problem of obtaining shape from shading would be solved since Horn's method could be used to trace out an arbitrary base characteristic, independent of the object in view. On the other hand, if constraints on the object surface $z = f(x,y)$ can be expressed in terms of properties of H, then these constraints can be applied to the set of possible solutions to the basic image-forming equation $I(x,y) = R(p,q)$. The Hessian matrix H is the generalization to $\mathbb{R}^n$ of the concept of curvature of a function of a real variable. Here, it is established that the corresponding positive (negative) definiteness of the Hessian is the generalization of positive (negative) curvature.

Definition. Let A be a real n x n symmetric matrix. Then A is called *positive semidefinite* if $x^TAx \geq 0$ for all $x \in \mathbb{R}^n$. A is called *positive definite* if $x^TAx > 0$ for all *nonzero* $x \in \mathbb{R}^n$. *Similarly*, A is called *negative semidefinite* if $x^TAx \leq 0$ for all $x \in \mathbb{R}^n$ and A is called *negative definite* if $x^TAx < 0$ for all nonzero $x \in \mathbb{R}^n$.

H is a real symmetric matrix (Corollary A.1). Therefore, it is appropriate to examine conditions under which H is positive (negative) definite. The positive (negative) definiteness of the Hessian matrix H is related to the convexity (concavity) of the corresponding (hyper)surface $z = f(x)$. The following definitions and theorems establish the required results:

Definition. A real-valued function $f(x)$ defined on a convex set $\Gamma \subset \mathbb{R}^n$ is said to be *convex* if, for every $x_1, x_2 \in \Gamma$ and every λ, $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

If, for every $0 < \lambda < 1$ and $x_1 \neq x_2$

$$f(\lambda x_1 + (1 - \lambda)x_2) < \lambda f(x_1) + (1 - \lambda)f(x_2),$$

then $f(x)$ is said to be *strictly convex*.
Similarly, a real-valued function $f(x)$ defined on a convex set $\Gamma \subset \mathbb{R}^n$ is said to be *concave* if $-f(x)$ is convex and *strictly concave* if $-f(x)$ is strictly convex.

> Theorem A.2 Let $f(x)$ be a real-valued function defined on an open convex set $\Gamma \subset \mathbb{R}^n$ and let the (hyper)surface described by the equation $z = f(x)$ be smooth over Γ. Then, $f(x)$ is convex on Γ if and only if $H = \nabla^2 f(x)$ is positive semidefinite on Γ. Similarly, $f(x)$ is concave on Γ if and only if $H = \nabla^2 f(x)$ is negative semidefinite on Γ.

Unfortunately, theorem A.2 does not extend to strictly convex and strictly concave functions by simply replacing the inequalities by strict inequalities. The extent by which it does extend is given by the following theorem.

> Theorem A.3 Let $f(x)$ be a real-valued function defined on an open convex set $\Gamma \subset \mathbb{R}^n$ and let the (hyper)surface described by the equation $z = f(x)$ be smooth over Γ. A sufficient but not necessary condition that $f(x)$ be strictly convex on Γ is that $H = \nabla^2 f(x)$ is positive definite on Γ. Similarly, a sufficient but not necessary condition that $f(x)$ be strictly concave on Γ is that $H = \nabla^2 f(x)$ is negative definite on Γ.

These results are used to show how convexity adds constraint. Suppose the surface $z = f(x,y)$ is convex. Multiplying the two equations $[I_x, I_y]^T = H\,[R_p, R_q]^T$ and $[dp, dq]^T = H\,[dx, dy]^T$ on the left by $[R_p, R_q]$ and $[dx, dy]$ respectively generates the two inequalities:

$$[R_p, R_q]\, H\, [R_p, R_q]^T = R_p I_x + R_q I_y \geq 0$$

$$[dx, dy]\, H\, [dx, dy]^T = dx\,dp + dy\,dq \geq 0$$

The first inequality $R_p I_x + R_q I_y \geq 0$ can be viewed as additional constraint on the contour in gradient space of possible solutions to the basic image-forming equation $I(x,y) = R(p,q)$. Suppose image point (x_0, y_0) has $I(x_0, y_0) = \alpha$. Figure A-2(a) shows the point (x_0, y_0) along with the corresponding image intensity contour $I(x,y) = \alpha$. The vector $[I_x, I_y]$ defines the direction normal to the contour $I(x,y) = \alpha$ at each point (x,y). Figure A-2(b) shows the reflectance map contour $R(p,q) = \alpha$. The vector $[R_p, R_q]$ defines the direction normal to the contour $R(p,q) = \alpha$ at each point (p,q).

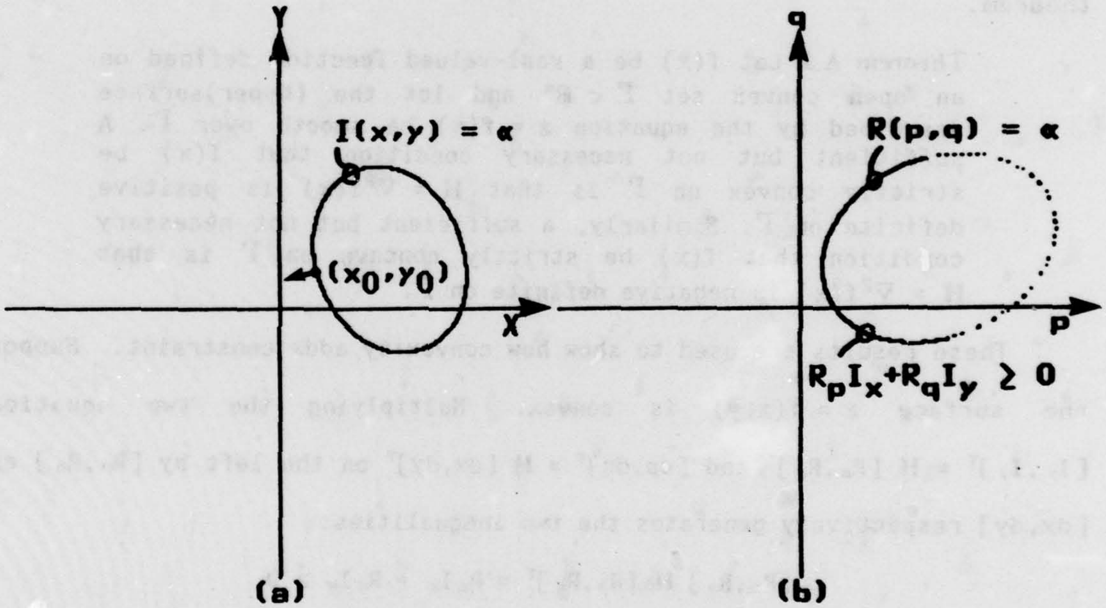


Figure A-2 The inequality $R_p I_x + R_q I_y \geq 0$ restricts the contour in gradient space that can correspond to a given image point.

$R_p I_x + R_q I_y \geq 0$ if and only if the angle between the normal $[I_x, I_y]$ and the normal $[R_p, R_q]$ is less than or equal to $90°$. Thus, if (x_0, y_0) lies on a section of surface known to be convex, points on the contour $R(p,q) = \alpha$ for which $R_p I_x + R_q I_y < 0$ (the dotted section of the contour $R(p,q) = \alpha$ of figure A-2(b)) can be excluded from the set of possible gradient points corresponding to image point (x_0, y_0).

The second inequality helps to constrain the movement in gradient space [dp,dq] that can correspond to a movement [dx,dy] in the image. Suppose image point (x_0, y_0) is known to correspond to the gradient point (p_0, q_0). The vector [dx,dy] defines the direction of movement in image space (figure A-3(a)). Now, $dxdp + dydq \geq 0$ if and only if the angle between [dx,dy] and the corresponding [dp,dq] is less than or equal to $90°$. Thus, if a movement is made in the direction [dx,dy] from the image point (x_0, y_0) on a section of surface known to be convex, then the corresponding movement in gradient space must take the point (p_0, q_0) into the region R of figure A-3(b).

This second result can be used to choose [dx,dy] in such a way as to guarantee that the view angle increases or that the direction of steepest descent increases. Figure A-4, illustrates what happens if [dx,dy] is chosen to be in the direction $[p_0, q_0]$. In this case, the line $dxdp + dydq = 0$ is the tangent line to the gradient space circle $p^2 + q^2 = p_0^2 + q_0^2$. All points in R now lie outside this circle so that the point (p_0, q_0) must move to a point of increasing view angle.

Figure A-5, illustrates what happens if [dx,dy] is chosen to be in the direction $[-q_0, p_0]$. In this case, the line $dxdp + dydq = 0$ is the line connecting (p_0, q_0) and the origin (0,0). All points in R lie above this line so that the point (p_0, q_0) must move to a point of increasing direction

of steepest descent.

A.3 A GEOMETRIC INTERPRETATION OF MULTIPLICATION BY THE IMAGE HESSIAN

A useful geometric interpretation of multiplication by a real symmetric matrix derives from the fact that any positive definite matrix can be used to define a norm. This provides a geometric interpretation of multiplication by a positive definite **H**. The interpretation can be extended to arbitrary **H**.

> Definition. Let A be an n x n positive definite matrix and let $x \in \mathbb{R}^n$. Then, $\|x\|_A$ is called the *A-norm of* x where
> $$\|x\|_A^2 = x^T A x$$

The key observation is that an A-norm is like the standard Euclidean norm except that it can apply different weights to the components of x in different directions. These weights and directions are determined by the eigenvalues and eigenvectors of A. The following three theorems formalize this observation.

> Theorem A.4 Let A be a real symmetric n x n matrix. Then all the eigenvalues of A are real and there exist n mutually orthogonal eigenvectors corresponding to each of the (not necessarily distinct) eigenvalues of A.

> Theorem A.5 Let A be positive definite. Then all the eigenvalues of A are positive.

> Theorem A.6 Let A be a positive definite n x n matrix. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the n positive but not necessarily distinct eigenvalues of A and let $\omega_1, \omega_2, \ldots, \omega_n$ be a corresponding set of n mutually orthogonal eigenvectors. (Without loss of generality, choose each ω_i to be a unit vector and order the ω_i so that the vectors $\omega_1, \omega_2, \ldots, \omega_n$ form a right-handed coordinate system.) Then, using $\omega_1, \omega_2, \ldots, \omega_n$ as a set of basis vectors, any n-vector x can be rewritten in the form:
> $$x = y_1\omega_1 + y_2\omega_2 + \ldots + y_n\omega_n$$
> Then:
> $$\|x\|_A^2 = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \ldots + \lambda_n y_n^2$$

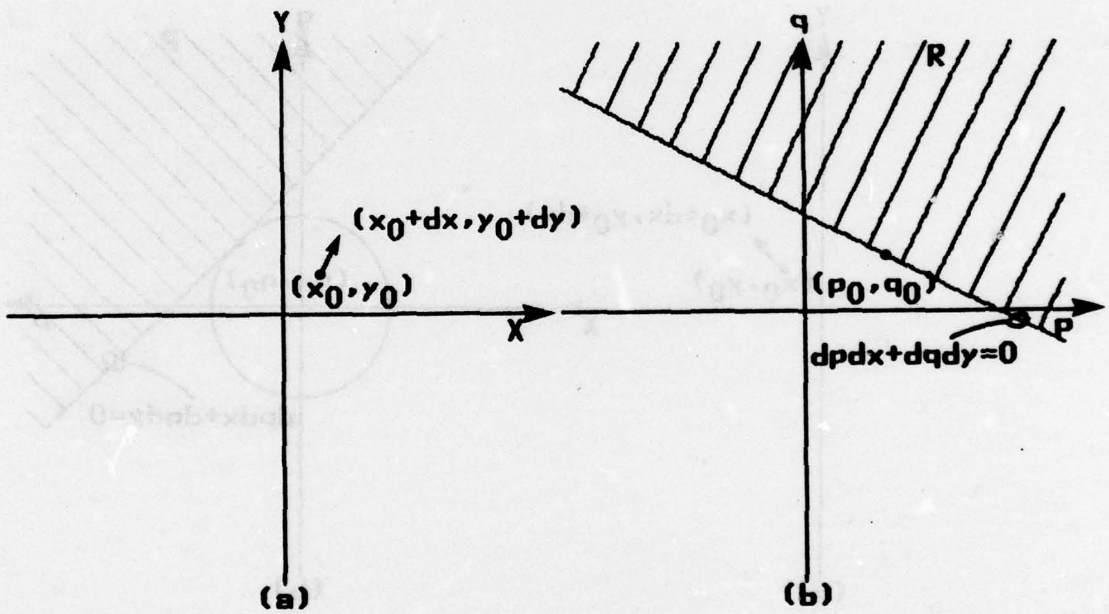


Figure A-3 The inequality dpdx + dpdy ≥ 0 restricts the movement in gradient space that can correspond to a given movement in image space.

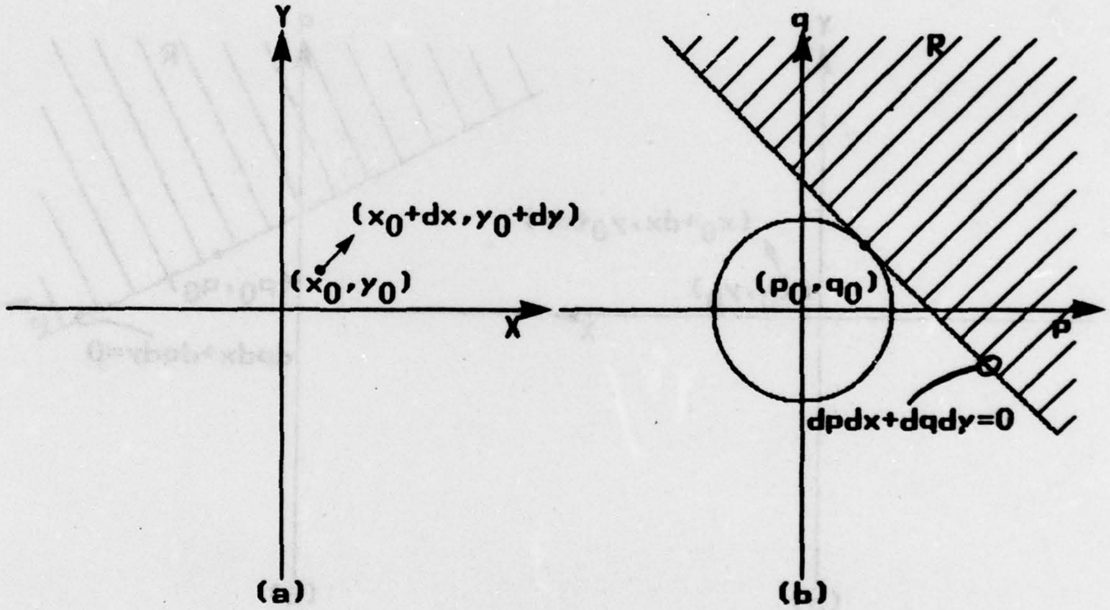


Figure A-4 Surface convexity can be used to choose a movement [dx,dy] in the image such that the corresponding movement [dp,dq] in gradient space increases the view angle e.

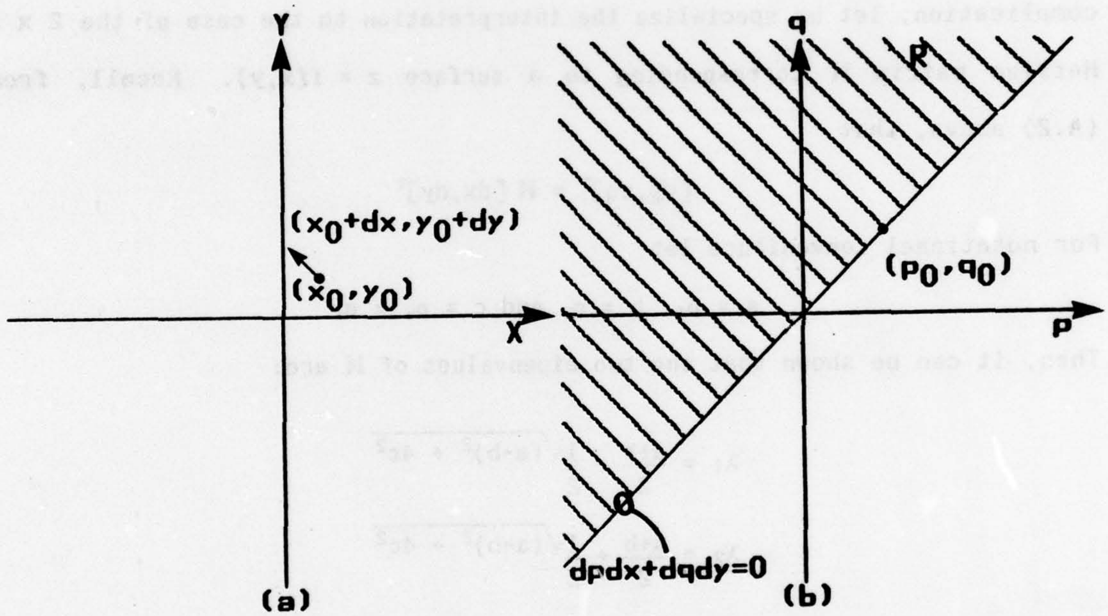


Figure A-5 Surface convexity can be used to choose a movement [dx,dy] in the image such that the corresponding movement [dp,dq] in gradient space increases the direction of steepest descent.

Theorem A.6 follows from the fact that any n x n symmetric matrix A is orthogonally (rotationally) similar to an n x n diagonal matrix whose diagonal elements are the eigenvalues of A. Knowing that all the eigenvalues of a positive definite matrix are positive guarantees that the quadratic form $x^T A x$ satisfies the requirements of a metric norm. Theorem A.6 provides a useful geometric interpretation for multiplication of a vector by a positive definite matrix. To avoid unnecessary complication, let us specialize the interpretation to the case of the 2 x 2 Hessian matrix H corresponding to a surface $z = f(x,y)$. Recall, from (A.2) above, that

$$[dp,dq]^T = H\,[dx,dy]^T$$

For notational convenience let

$$a = p_x,\ b = q_y \text{ and } c = p_y = q_x$$

Then, it can be shown that the two eigenvalues of H are:

$$\lambda_1 = \frac{a+b}{2} - \frac{1}{2}\sqrt{(a-b)^2 + 4c^2}$$

$$\lambda_2 = \frac{a+b}{2} + \frac{1}{2}\sqrt{(a-b)^2 + 4c^2}$$

and that the corresponding (unit) eigenvectors are:

$$\omega_1 = [\cos(\theta), -\sin(\theta)]$$

$$\omega_2 = [\sin(\theta), \cos(\theta)]$$

where

$$\tan(2\theta) = \frac{2c}{b-a}$$

Multiplication by H can then be interpreted as follows: [dp,dq] is the vector sum of the projection of [dx,dy] in each of the eigenvector directions ω_1 and ω_2 where each projection is scaled by the corresponding eigenvalues λ_1 and λ_2. That is, $[dp,dq] = \lambda_1(v^T\omega_1)\omega_1 + \lambda_2(v^T\omega_2)\omega_2$ where

$v^T = [dx,dy]$.

Consider moving a small distance ds in the image. Consider the family of all [dx,dy] such that $ds = \sqrt{dx^2 + dy^2}$. If H is positive definite at an image point (x_0,y_0) known to correspond to the gradient point (p_0,q_0), then multiplication by H induces a 1-1 continuous mapping of the image space circle centered at (x_0,y_0) and with radius ds onto the gradient space ellipse centered at (p_0,q_0) and with axes $\lambda_1 ds$ and $\lambda_2 ds$ in the directions ω_1 and ω_2 respectively. Figure A-6 illustrates this result, with image space coordinates and gradient space coordinates superimposed. The fact to be exploited is that the mapping from this image space circle to the corresponding gradient space ellipse is continuous and 1-1. Multiplication by a positive definite H is monotonic in the following sense: multiplication by a positive definite H preserves the ordering in angular position of the [dp,dq]'s corresponding to given [dx,dy]'s. Let us make this more precise.

> **Definition.** Suppose two nonzero vectors $x = [x_1,x_2]$ and $y = [y_1,y_2]$ are described in a right-handed (positive) coordinate system. Then x is said to be *(strictly) less in angular position* than y if the angle required to align x with y by rotating x in a counter-clockwise direction is (strictly) less than the angle required to align x with y by rotating x in a clockwise direction. Similarly, x is said to be *(strictly) greater in angular position* than y if the angle required to align x with y by rotating x in a counter-clockwise direction is (strictly) greater than the angle required to align x with y by rotating x in an clockwise direction.

> **Definition.** Let $z = f(x,y)$ be the equation describing a smooth surface and let λ_1 and λ_2 be the two eigenvalues of the corresponding Hessian matrix H at image point (x_0,y_0). The surface $z = f(x,y)$ is said to be *planar at* (x_0,y_0) if and only if $\lambda_1 = \lambda_2 = 0$. The surface $z = f(x,y)$ is said to be *singly curved at* (x_0,y_0) if and only if one (but not both) of λ_1 and λ_2 is equal to zero. The surface $z = f(x,y)$ is said to be *doubly curved at* (x_0,y_0) if and only if both λ_1 and λ_2 are not equal to zero.

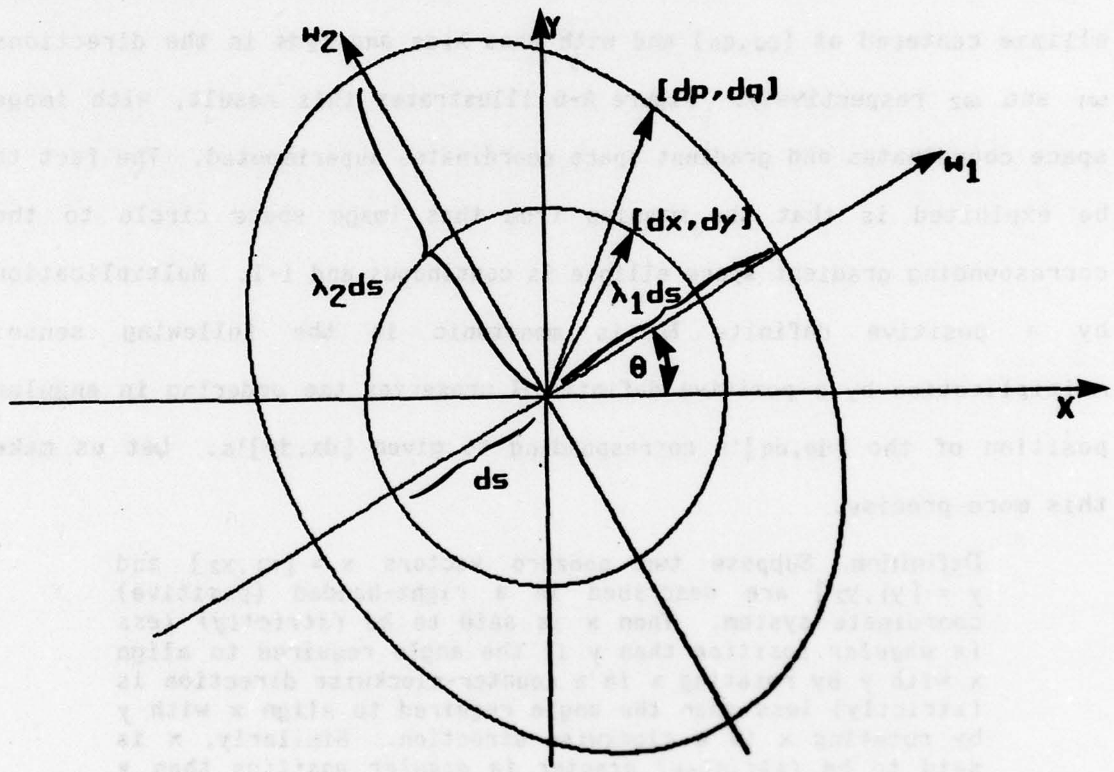


Figure A-6 Multiplication by the image Hessian H induces a 1-1 mapping between a circle of radius ds in the image and an ellipse in gradient space. The major and minor axes of this ellipse are determined by the eigenvalues and eigenvectors of H.

Theorem A.7 Let $z = f(x,y)$ be the equation describing a smooth surface. Then $z = f(x,y)$ is doubly curved at (x_0,y_0) if and only if the corresponding Hessian matrix H is nonsingular at (x_0,y_0).

Theorem A.8 Let $z = f(x,y)$ be the equation describing a smooth surface and suppose $z = f(x,y)$ is doubly curved at image point (x_0,y_0). Let H be the corresponding Hessian matrix at (x_0,y_0) having nonzero eigenvalues λ_1 and λ_2.

(i) If H is positive or negative definite (λ_1 and λ_2 have the same sign), then multiplication by H preserves the ordering of angular positions of $[dp,dq]$'s with respect to the corresponding $[dx,dy]$'s. That is, if $[dp_1,dq_1]^T = H\,[dx_1,dy_1]^T$, $[dp_2,dq_2]^T = H\,[dx_2,dy_2]^T$ and $[dx_1,dy_1]$ is (strictly) less in angular position than $[dx_2,dy_2]$ then $[dp_1,dq_1]$ is (strictly) less in angular position than $[dp_2,dq_2]$.

(ii) If H is neither positive nor negative definite (λ_1 and λ_2 have opposite sign), then multiplication by H reverses the ordering of angular positions of $[dp,dq]$'s with respect to the corresponding $[dx,dy]$'s. That is, if $[dp_1,dq_1]^T = H\,[dx_1,dy_1]^T$, $[dp_2,dq_2]^T = H\,[dx_2,dy_2]^T$ and $[dx_1,dy_1]$ is (strictly) less in angular position than $[dx_2,dy_2]$ then $[dp_1,dq_1]$ is (strictly) greater in angular position than $[dp_2,dq_2]$.

Theorem A.8 requires that $z = f(x,y)$ be doubly curved in order to guarantee that multiplication by H is a 1-1 mapping. However, if H is nonsingular, then theorem A.8 gives precise conditions with which to order the angular changes to position in gradient space corresponding to given movements in image space. If H is either positive or negative definite then the mapping of image space circle to gradient space ellipse goes from a right-handed (positive) coordinate representation to a right-handed (positive) coordinate representation so that the ordering of angular positions is preserved. If H is neither positive nor negative definite then the mapping of image space circle to gradient space ellipse goes from a right-handed (positive) coordinate representation to a left-handed (negative) coordinate representation so that the ordering of angular positions is reversed.

A.4 THE IMAGING MATHEMATICS OF A SPHERE

Expressions can be developed for the surface orientation and the image Hessian matrix **H** for an image of a sphere. The purpose is not to perform an exercise in differential calculus. Rather, the analytic results developed here can be used to illustrate the connection between a viewer-centered definition of the image Hessian matrix **H** and the more traditional object-centered definition of curvature.

Consider a sphere of radius r centered at the object space origin. The surface of the sphere is described implicitly by the equation:

$$F(x,y,z) = x^2 + y^2 + z^2 - r^2 = 0$$

This equation gives no indication of which points on the surface project to points in the image nor which points on the surface are hidden from view. It is an object-centered representation of the surface. For the work here, an explicit representation is required. For a sphere, this explicit representation is given by the equation:

$$z = f(x,y) = -\sqrt{r^2 - x^2 - y^2}$$

An explicit representation of the form $z = f(x,y)$ is a viewer-centered representation of the object surface. Recall, from figure 2-3, that the viewer is looking along the positive z-axis so that the points on the sphere actually in view correspond to negative values of z as indicated above.

The gradient coordinates p and q are determined by differentiating $f(x,y)$ with respect to x and y. One finds:

$$p = \frac{\partial f(x,y)}{\partial x} = \frac{-x}{z}$$

$$q = \frac{\partial f(x,y)}{\partial y} = \frac{-y}{z}$$

Taking second partial derivatives of $f(x,y)$ with respect to x and y, one

finds:

$$p_x = \frac{\partial^2 f(x,y)}{\partial x^2} = -\frac{(r^2-y^2)}{z^3}$$

$$q_y = \frac{\partial^2 f(x,y)}{\partial y^2} = -\frac{(r^2-x^2)}{z^3}$$

$$p_y = q_x = \frac{\partial^2 f(x,y)}{\partial x \partial y} = -\frac{xy}{z^3}$$

Thus, the Hessian matrix H is given by:

$$H = -1/z^3 \begin{bmatrix} r^2-y^2 & xy \\ xy & r^2-x^2 \end{bmatrix}$$

The eigenvalues of the Hessian matrix H are given by:

$$\lambda_1 = \frac{-1}{z}$$

$$\lambda_2 = \frac{-r^2}{z^3}$$

with corresponding (unit) eigenvectors:

$$\omega_1 = [\sin(\theta), -\cos(\theta)]$$
$$\omega_2 = [\cos(\theta), \sin(\theta)]$$

where

$$\tan(\theta) = \frac{y}{x}$$

First, consider the special case $x = 0$ $y = 0$. Then $z = f(0,0) = -r$, $p = 0$, $q = 0$ and the Hessian matrix H becomes:

$$H = \begin{bmatrix} 1/r & 0 \\ 0 & 1/r \end{bmatrix}$$

This is as expected. The curvature in each of the two principal directions of movement is the same, and is equal to one over the radius of the sphere. This is actually true for any point on the surface of a sphere.

(Conversely, a sphere is the only surface which has constant nonzero curvature in all directions and at all points.) But, this notion of surface curvature corresponds to an object-centered representation. The image Hessian matrix **H** is not constant for all (x,y).

A.5 RELATING THE IMAGE HESSIAN TO SURFACE CURVATURE

The way the image Hessian matrix **H** has been defined in this report corresponds to a viewer-centered representation of curvature. The Hessian matrix **H** relates movement in the image to changes in local surface orientation and not movement on the object surface to changes in local surface orientation. The case x = 0 y = 0, cited for the sphere above, is the unique situation for which the object-centered definition of curvature and the *viewer-centered definition of curvature* coincide. Here, the Hessian matrix **H** intuitively captures what one would expect from a "curvature" matrix. In a viewer-centered representation, the image Hessian matrix **H** of a sphere is not constant. The dependence of **H** on x and y captures the variation in apparent curvature when the surface is viewed obliquely.

This is made clear if the expression for the image Hessian **H** of a sphere is rewritten in terms of gradient coordinates p and q. One finds:

$$\mathbf{H} = 1/\cos(e)\begin{bmatrix} p^2+1 & pq \\ pq & q^2+1 \end{bmatrix}\begin{bmatrix} 1/r & 0 \\ 0 & 1/r \end{bmatrix} \tag{A.3}$$

where e is the view angle. That is:

$$\cos(e) = \frac{1}{\sqrt{1+p^2+q^2}}$$

The following two theorems provide an interpretation for (A.3) above:

> **Theorem A.9** Let $z = f(x,y)$ be the equation describing a smooth surface. Let dA be a differential element of area on the surface. Then,
>
> $$dA = \frac{1}{\cos(e)} dxdy$$
>
> where $dxdy$ is the corresponding differential element of area in the image and e is the view angle subtended by the surface element dA. Thus, the surface area corresponding to a region R in the image is given by:
>
> $$A = \iint_R \sec(e)\, dxdy$$

> **Theorem A.10** Let $z = f(x,y)$ be the equation describing a smooth surface. Let ds be the differential of the arc on the surface corresponding to a movement $[dx,dy]$ in the image. Then, ds is given by the A-norm of $[dx,dy]$. That is,
>
> $$ds = \|[dx,dy]\|_A = \sqrt{[dx,dy]A[dx,dy]^T}$$
>
> where the matrix A is given by:
>
> $$A = \begin{bmatrix} p^2+1 & pq \\ pq & q^2+1 \end{bmatrix}$$

Theorem A.9 allows one to interpret multiplication by $1/\cos(e)$ in (A.3) as compensation for the foreshortening of area due to the oblique view corresponding to a gradient (p,q). Theorem A.10 allows one to interpret multiplication by A in (A.3) as compensation for the foreshortening of path length due to the oblique view corresponding to a gradient (p,q). Finding the area of surface corresponding to a given region of image depends only on the magnitude of the gradient at each image point in the region. On the other hand, finding the path length along the surface corresponding to a given curve in the image depends on both the magnitude and angular position of the gradient at each image point on the curve.

This dependence of path length on the gradient (p,q) can be made more explicit by, once again, examining the eigenvalue and eigenvector structure of A. The matrix A is positive definite with eigenvalues:

$$\lambda_1 = 1$$

$$\lambda_2 = \frac{1}{\cos^2(e)}$$

and corresponding (unit) eigenvectors:

$$\omega_1 = [\sin(\theta), -\cos(\theta)]$$

$$\omega_2 = [\cos(\theta), \sin(\theta)]$$

where

$$\tan(\theta) = \frac{q}{p}$$

The component of the differential ds of the arc on the surface in the direction of steepest descent is foreshortened by the factor cos(e) while the component of the differential ds in the direction of the contour of constant $z = f(x,y)$ is unchanged.

Since A is positive definite, it is also invertible and its inverse A^{-1} is positive definite. Knowing the gradient point (p,q) and the Hessian matrix H at an image point (x,y) allows one to determine the magnitude and direction of the object-centered principal radii of curvature of the surface $z = f(x,y)$. Suppose that k_1 and k_2 are the two eigenvalues and ω_1 and ω_2 are the corresponding unit eigenvectors of the matrix C where:

$$C = \cos(e) A^{-1} H = \cos^3(e) \begin{bmatrix} q^2+1 & -pq \\ -pq & p^2+1 \end{bmatrix} H$$

Then, $r_1 = 1/k_1$ and $r_2 = 1/k_2$ are the two principal radii of curvature of the surface $z = f(x,y)$ oriented respectively in the directions defined by ω_1 and ω_2.

Finally, since $\cos(e) > 0$ and A^{-1} is positive definite, one observes that the number of nonzero eigenvalues of H is equal to the number of nonzero eigenvalues of $C = \cos(e)A^{-1}H$. An eigenvalue λ_i of H is zero (nonzero) precisely as the corresponding principal radius of curvature r_i is infinite (finite). Thus, the definitions of a surface $z = f(x,y)$ being planar at (x_0,y_0), singly curved at (x_0,y_0) and doubly curved at (x_0,y_0) given in terms of the viewer-centered Hessian matrix H are equivalent to the more standard definitions given in terms of the object-centered principal radii of curvature $r_1 = 1/k_1$ and $r_2 = 1/k_2$.

APPENDIX B: CATALOGING CASTING DEFECTS

The initial goal in exploring casting inspection as a suitable area of application for machine vision was to develop a simple catalog of defects in metal castings. The first observation to be made is that the phrase "defects in metal castings" encompasses a broad spectrum of possible issues. The defects asssociated with a particular casting process fall roughly into one of three categories:

1. INHERENT DEFECTS
   - defects introduced during the preparation of the metal alloy or other raw materials

2. PROCESSING DEFECTS
   - defects introduced during the casting process

3. SERVICE DEFECTS
   - defects introduced during the operating cycle of the casting

Here, only defects introduced by the casting process itself are considered. It is appropriate to further split such processing defects into three categories:

1. SHAPE DEFECTS
   - defects that affect the overall dimensional accuracy of the part

2. SURFACE DEFECTS
   - defects that manifest themselves as local surface properties of the part

3. STRUCTURE DEFECTS
   - defects that affect the mechanical response of the part

These three categories of processing defects are not mutually exclusive. There is considerable overlap between categories. The kind of inspection considered in this work relates primarily to SURFACE defects. SHAPE defects related to the accurate verification of part dimensions are excluded. STRUCTURE defects which can be found only by destructive

testing, special test equipment or by imaging rays which penetrate the object surface (eg. X-rays, ultra-sonic rays) are also excluded. As the catalog below demonstrates, many defects which affect the structural properties of a casting are nonetheless manifest as surface properties. These are not excluded.

Before presenting the catalog, a slight disclaimer is in order: this catalog of casting defects really should take into account the particular alloy used, the particular casting technique chosen and the particular part geometry. Fortunately, however, each alloy, each casting technique and each part geometry share the same broad categories of defects. The difference lies in the relative rate of occurrence of the various kinds of defects and, to a lesser extent, in the way in which these defects manifest themselves.

With that disclaimer, here is the catalog of defects:

COLD SHUT (FLOW MARK)

DESCRIPTION:

A defect that affects both the SURFACE and the STRUCTURE of a casting. Broadly speaking, a cold shut defect arises when molten metal does not weld together properly. There are two types of cold shut defect. One type occurs at the interface of two streams of molten metal whose temperature differential prevents them from welding together properly. This is manifest as a lapping or layering on the surface of the casting. The other type results from loose droplets of molten metal entering the mold cavity ahead of the main stream of metal. These solidify and become partially embedded in the surface of the casting.

CAUSES

DESIGN:

- portions of the mold too cold

PRODUCTION:
- pouring/injection rate too slow
- mold temperature too low
- metal temperature too low
- dirty metal (gating blockages)

CRACKS

DESCRIPTION:

A defect that affects both the SURFACE and the STRUCTURE of a casting. A crack is a defect that occurs during contraction shrinkage following solidification. Localized stresses are set up within a casting immediately following solidification, especially at junctions of restraining ribs, right angle intersections, and junctions between thick and thin sections. These stresses can produce cracks as the casting cools to room temperature. Because such cracks occur after solidification, there is little chance for oxidization and they are typically very clean.

CAUSES

DESIGN:
- uneven or too rapid cooling
- excessive mold rigidity restraining normal contraction of the metal

PRODUCTION:
- mechanical jarring during removal from the mold

HOT TEARS

DESCRIPTION:

A defect that affects both the SURFACE and the STRUCTURE of a casting. A hot tear is a cracklike defect that occurrs during solidification. As the molten metal touches the comparatively cold mold surface, it soldifies as a skin and starts contracting ahead of the remainder of the casting. This skin is placed in tension and may tear if it becomes overstressed. These tears are characterized by having a heavily oxidized surface while CRACKS (see above), which occur after solidification, are relatively clean.

CAUSES
DESIGN:
- uneven or too rapid cooling
- excessive mold rigidity restraining normal contraction of the metal
- poor pattern design
  too small a radius of curvature at a section boundary
  too high a ratio of areas between sections joined in a T

INCLUSIONS

DESCRIPTION:

Inclusions are defects that arise due to foreign material trapped in a casting. Inclusions affect both the SURFACE and the STRUCTURE of a casting. Broadly speaking, inclusions are classified according to their origin.

INCLUSIONS OF METALLIC ORIGIN (DROSS/SLAG)

Improper melting and pouring practice may cause metallic inclusions in the casting. The formation of oxides, slag and other metallic waste material is an inherent part of the melting process. This slag material is lighter than the molten metal and floats to the surface. Careful design of the metal feeding system attempts to take advantage of this fact to prevent slag material from entering the mold cavity.

CAUSES
DESIGN:
- faulty gating design

PRODUCTION:
- poor quality control on raw materials
- superheating of metal in melting furnace
- melting cycle too long
- turbulent flow of metal

INCLUSIONS OF NONMETALLIC ORIGIN

Some of the possible sources of nonmetallic inclusions are:

Mold and core material
  Extraneous mold and core material may remain loose in the mold cavity or may be generated by the erosive action of the incoming molten metal.

Pattern material

In investment casting, residue from wax or plastic patterns may remain in the mold cavity and contaminate the casting.

Crucible and furnace lining material

Lining breakdown in crucibles and melting furnaces adds contaminants to the molten metal.

(PARTING LINE) MISMATCH

DESCRIPTION:

A defect that alters both the SURFACE and the SHAPE of a casting. (Parting line) mismatch occurs in casting techniques that employ a two-piece mold. Misalignment between the mold halves induces a step shift at the parting line of the casting. This results in a surface irregularity and loss of dimensional accuracy. The tolerable degree of parting line mismatch is generally specified in the mechanical drawing of a casting.

CAUSES

DESIGN:

- misalignment of pins and receptacles of the mold halves

PRODUCTION:

- play (due to wear over time) developing between pins and receptacles of mold halves

MISRUN

DESCRIPTION:

A SHAPE defect due to the incomplete filling of the mold cavity. Misruns occur when an advancing stream of molten metal lacks sufficient force to overcome back pressure generated in the mold. (Typically, misruns occur in isolated thin sections of a casting.)

CAUSES

DESIGN:

- faulty gating design (need for additional vents, overflows or altered direction of metal flow)
- low mold permeability
- incorrect mold/metal temperatures

PRODUCTION:

- dirty metal (gating blockages)
- failure of mold reaction inhibitors (excessive evolution of gases)
- overlubrication (inhibits permeability)
- pouring/injection rate too slow
- inadequate quantity of molten metal in shot well
- mold temperature too low
- metal temperature too low

POROSITY

DESCRIPTION:

There is some confusion about the use of the term porosity. When not distinctly referring to shrinkage porosity, porosity generally implies bubbles of gas entrapped in the metal during solidification. Many of these bubbles remain internal to the casting and thus represent STRUCTURE defects (internal porosity). Certain others reach the surface during solidification and thus represent SURFACE defects (surface porosity). Large surface porosity defects are commonly referred to as BLOWHOLES. Small surface porosity defects are commonly referred to as PINHOLES. There are three main sources of entrapped gas in a casting:

1. Gas evolved within the molten metal itself. (Gas solubility in a metal decreases as temperature decreases.)

2. Air trapped in mold cavity

3. Gas evolved due to chemical reactions between the casting metal and the mold or core material

CAUSES

DESIGN:

- faulty gating design
- improper mold reaction inhibitors
- improper innoculants
- low mold permeability
- incorrect mold/metal temperatures
- poor metal handling

PRODUCTION:
- overlubrication
- pouring/injection rate too slow
- turbulent flow of metal
  - pouring/injection rate too high
  - partial blockage in gating system
- mold temperature too high
- metal temperature too high
- contamination of raw materials

SHRINKS

(SHRINK CAVITIES, SHRINKAGE VOIDS, SHRINKAGE POROSITY, PIPE)

DESCRIPTION:

In cooling from a molten state to room temperature, metal goes through three stages of shrinkage:

- volumetric shrinkage as a liquid
- solidification shrinkage during conversion to a solid
- contraction shrinkage as a solid cooling to room temperature

Shrinkage defects (commonly referred to a shrink cavities, shrinkage voids, shrinkage porosity or pipe) are STRUCTURE defects. They arise due to metal shrinkage within a shell of already solidified metal that is not compensated for by a continued inflow of molten metal. Shrinkage defects typically occur near the center of large, heavy sections of a casting.

CAUSES

DESIGN:
- faulty gating design
- poor control of direction of solidification

PRODUCTION:
- pouring/injection rate too slow
- mold temperature to high
- metal temperature too high
- carbon equivalent of metal too low

MISCELLANEOUS METAL/MOLD INTERACTION DEFECTS

In the discussion of green sand mold casting, we have already seen a table of miscellaneous surface defects that result from a poor metal-to-mold interface. In permanent mold casting techniques, there are similar defects.

SOLDERING:

Soldering is a SURFACE defect in permanent mold castings due to the adhering of metal to the mold surface. Soldering results in pimples or torn skin on the surface of the casting. It arises when the mold surface has become pitted or when there is inadequate lubrication between mold and metal.

STAINS:

Excessive lubrication of the mold surface can result in stains on the surface of the casting.